Simulation and Knowledge of Action

Edited by
Jérôme Dokic and
Joëlle Proust

Advances in Consciousness Research

Simulation and Knowledge of Action

Advances in Consciousness Research

Advances in Consciousness Research provides a forum for scholars from different scientific disciplines and fields of knowledge who study consciousness in its multifaceted aspects. Thus the Series will include (but not be limited to) the various areas of cognitive science, including cognitive psychology, linguistics, brain science and philosophy. The orientation of the Series is toward developing new interdisciplinary and integrative approaches for the investigation, description and theory of consciousness, as well as the practical consequences of this research for the individual and society.

Series B: Research in progress. Experimental, descriptive and clinical research in consciousness.

Volume 45

Simulation and Knowledge of Action
Edited by Jérôme Dokic and Joëlle Proust

Simulation and Knowledge of Action

Edited by

Jérôme Dokic and
Joëlle Proust
Institut Jean-Nicod, CNRS, Paris

John Benjamins Publishing Company
Amsterdam/Philadelphia

Library of Congress Cataloging-in-Publication Data

Simulation and Knowledge of Action / edited by Jérôme Dokic and Joëlle Proust.
p. cm. (Advances in Consciousness Research, ISSN 1381–589X ; v. 45)
Includes bibliographical references and indexes.
1. Philosophy of mind in children. I. Dokic, Jerôme, 1965- II. Proust, Joëlle. III. Series.

BF723.P48 S56 2002
128'.2-dc21 2002074693
ISBN 90 272 5169 X (Eur.) / 1 58811 266 7 (US) (Hb; alk. paper)
ISBN 90 272 5170 3 (Eur.) / 1 58811 267 5 (US) (Pb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Introduction

Jérôme Dokic and Joëlle Proust
Institut Jean-Nicod, CNRS, Paris

Is our understanding of the mind to be explained in terms of our mastering a theory, or in terms of a general capacity for simulating other agents? The first type of theory, called the "Theory Theory", claims that children learn how to understand and predict other people's psychological states in a way that closely parallels how they learn physical and biological facts. They posit unobservable theoretical entities (mainly beliefs and desires) in order to explain observed facts (mainly behaviour), master the inferential relations between them in a given domain, and learn how to apply the resulting concepts to new cases within the domain. There has been no consensus among the developmental psychologists and philosophers involved in this debate as to the role respectively attributed to experience and to innate constraints on mentalising. Some of the main proponents of the idea of a Theory Theory of mind have hypothesised that children rely both on experience and on their general reasoning capacities to produce folk-psychological explanations.[1] Others have claimed that theory of mind originates in a domain-specific module that comes to fruition at around 2 years of age.[2] This module is taken to consist in a meta-representational knowledge structure that allows children to represent facts about propositional attitudes by "decoupling" a primary representation, with its usual reference, from a secondary representation, whose reference depends on its specific role in the context of a pretence, belief, etc.

While the Theory Theory of mental concept acquisition takes its inspiration from an analogy with the case of linguistics (in its modular version) or with general scientific procedures (in its non-modular version), the second type of theory, called "Simulation Theory", arises from two independent sets of considerations. One is the kind of philosophical consideration associated with the principle of charity; to understand others necessarily presupposes viewing them as thinking and rationally evaluating situations as I do. Simulating oth-

ers – i.e. using one's own evaluation and reasoning mechanisms as a model for theirs – allows us to put them into the required perspective as rational agents. Perhaps this is the only possible way to do this if, as some philosophers have claimed, rationality is not codifiable,[3] so that there is a sense in which we cannot devise a completely detached or perspective-free account of rationality and propositional attitudes. The relevant perspective is also empirically adequate: it maximises our chances of understanding all the emotional and factual dimensions of others' decisions, independently of the mastery of any specific psychological theory.[4]

The second kind of consideration relies on evidence from the cognitive sciences. It consists in registering the presence of simulatory mechanisms in a number of different fields of mental activity, not all of which have to do with the understanding of other minds.[5] A subject who is engaging in simulation may "replay" her own past experience in a direct, plain manner, in order to extract from it pleasurable, motivational or strictly informational properties. Simulatory activities such as engaging in perceptual imagery, activating episodic memories, training mentally in sports, exemplify this kind of function. On the other hand, simulation may also be put to a more distinctively innovative use, and allow experienced features to be combined in new ways. This is the case when simulating is preparatory to decision making, when it is used in pretend games, in fiction, and in understanding others.

Simulating in this sense consists in using information in an immersed, recreative or productive way. By "immersed", at least two things are meant. First, the representations used in simulation include phenomenological and spatial properties organised from a perceiver's point of view; simulation is essentially perspectival. Second, simulation is immersed in the sense that it can be entirely practical; the subject does not reflect on the processes of simulation themselves, which are normally cognitively impenetrable. For instance, replaying one's own past experience may be triggered by contextual cues, independently of the subject's intentions, and even unconsciously. The contrast between cognitive immersion and detachment is sometimes interpreted as a distinction between two ways – practical and theoretical – of *representing* reality.[6] However, in basic cases of simulation, it is not even clear that a distinctive level of (personal) *content* is involved. Simulation as a practical ability to anticipate behaviour may involve mainly subpersonal mechanisms.[7] This is also the case for all the resonance phenomena that have been observed in the brains of people watching an action executed by someone else, or of someone looking at another subject experiencing pain.[8]

Note that the two kinds of considerations offered in favour of Simulation Theory may not, in the last analysis, peacefully coexist in an inclusive explanatory framework. For the first kind of argument involves *a priori* considerations about rational evaluation issued at the first-person level, whereas the second kind includes facts that belong to a subpersonal level where information-processing operations are executed independently of what a subject may know or intend to do.

Simulation Theory and Theory Theory differ on a variety of accounts, which we shall not exhaustively rehearse in this introduction; they will be scrutinised in the various contributions of this volume. A first major difference between a simulatory account and an interpretation in terms of theory possession is epistemological: the resources used in simulation include all the facts of the world, whereas Theory Theory concentrates on mental inferences. A second contrast is that Simulation Theory and Theory Theory do not provide the same kind of output. Simulating triggers pretend-mental states (pretend-beliefs, pretend-desires, pretend-emotions) and evaluations relative to a concrete, individual situation. This ability is related to empathy; but it does not *per se* explain how children acquire mental concepts on the basis of their capacity to pretend to be in someone else's state. Some would claim that the very ability to attempt to simulate another subject presupposes – rather than mediates – such mental concept possession.[9] Theory Theory, on the other hand, aims at providing an explanation for mental-concept acquisition and reasoning; but it has nothing to say about how a child can apply mental concepts to individual cases.

The formerly hot debate between proponents of Theory Theory and Simulation Theory has gradually evolved into an attempt to develop hybrid theories, paying their respective dues to simulatory procedures and conceptual knowledge.[10] For instance, even if simulation on its own cannot explain our mastery of mental concepts, it may be that a complete account of our possession of these concepts must invoke simulative abilities. Such theoretical compromises may however be deemed to have come too early, i.e. even before conceptual analysis has come to full maturity. Two classes of problems seem to prevent the debate between simulation and theory from getting off the ground. One class concerns the question of what *mental concept possession* involves: is it to be equated with an ability to use these concepts in explaining or predicting action in an explicit way? Or should one rather accept that concepts can be mastered and used in a tacit way? A subject might in this view possess mental concepts while being unable to apply them in explaining others' actions (or her own). A further point to be discussed is whether possession of a tacit theory of mind

is compatible with a child's failure to use mental concepts consistently in her decisions and utterances. Some authors indeed defend the view that a child may have mastered a mental knowledge structure while being prevented, on executive grounds, from expressing her knowledge in her behaviour.[11]

This first class of questions plagues both theories, although in different ways: Theory Theory can be seen as having to carry the burden of clarifying further what a "theory" means in the context of development. Simulation Theory, on the other hand, needs to bridge the gap between having pretend-mental states and categorising these mental states for the sake of psychological explanation.

A second class of difficulties concerns what it means exactly to *"put oneself into another's shoes"*. Full projection of one's own beliefs and desires to the individual to be simulated would result in errors of interpretation. Projection can be successful only if it manages to set aside one's mental states not shared by the target, while somehow taking into account states that are not the simulator's own. But how is this to be done without using mental concepts? If, as radical simulationism tells us, projecting consists essentially in recentering one's egocentric map on the individual to be simulated,[12] to whom does "I" refer in the course of the projection? How can a simulator distinguish herself from the person she is projecting to?

All these difficulties seem to point to the fact that Simulation Theory, as things stand, offers an explanation schema rather than a *bona fide* theory. More work, both empirical and conceptual, needs to be done, in order to give a more precise meaning to the very idea of "putting oneself in someone else's shoes", and to understand how mental concepts can be extracted from an ability to simulate other people or foreign situations. The collection of essays below represents an effort at revising and extending Simulation Theory, while also considering areas in which it could be submitted to experimental tests.

Alvin Goldman's chapter contrasts Theory Theory and Simulation Theory as empirical theories assessed by cognitive science. According to him, developmental evidence does not clearly favour Theory Theory. In particular, there is no definite evidence that there is a move from a non-representational to a representational understanding of the mind, mastered by age 4. For instance, a failure on false-belief tasks before that age might indicate an executive deficit rather than a conceptual or theoretical one. Similarly, the supposedly non-representational concept of desire that 2-year-olds possess, according to some versions of Theory Theory, is in fact representational, since these children understand the possibility of unfulfilled desires, which are, Goldman claims, analogous to false beliefs.

Moreover, the claim that first-person attributions are theory-mediated is not very plausible. Friends of Theory Theory have accordingly proposed a direct, non-inferential "detection account" of self-awareness. However, Goldman argues that such an account is incompatible with functionalism, the Theory Theory's bedfellow, since it is unclear how we are supposed to directly detect our own mental states if they are functional, dispositional properties.

Goldman himself favours Simulation Theory as the primitive, default method of attribution. Part of the originality of his defence of that theory is that simulation is viewed as belonging to a generic type of heuristic along with imitation, pretend play, drama, visual and motor imagery. So although there is no direct empirical evidence that pretend-beliefs and desires *resemble* the corresponding non-pretend states, indirect evidence for this claim can be gathered from other cognitive domains belonging to the same, re-creative type of heuristic. For instance, there *is* evidence that motor imagery shares important resources with actual action.

Goldman tentatively puts forward a "dual-representation" hypothesis about mental state concepts. At first, children have two sorts of representations, respectively about behavioural characteristics and about internal (and perhaps, but not necessarily phenomenological) characteristics, which they then learn to conceptualise as representations of one and the same mental state. Resonance phenomena (such as the ones involving Rizzolatti's "mirror-neurons"[13] might facilitate attributing the internal characteristics of mental states to others. The idea is that such phenomena may prompt the understanding that we can share with another subject a type of mental state both internally and behaviorally defined.

In his reply, Bill Child raises a number of worries about Goldman's account. First, he argues that we can rescue the notion of a non-representational concept of desire, and he suggests two ways of doing so. Second, he claims that Goldman's analogy between a false belief and an unfulfilled desire is not an exact one; what is most closely analogous to the idea of a false belief is in fact the idea of someone's desiring something undesirable (from one's point of view). So Goldman's arguments against a gradual development from a non-representational understanding of the mind to a representational understanding is flawed. Finally, Child thinks that Goldman's account still faces the conceptual problem of other minds, which is how someone can get a concept of experience that allows her to understand the thought that others have conscious experiences. He is sceptical about the centrality of mirror neurons and resonance phenomena in the explanation of our grasp of a concept of conscious experience equally applicable to others and to oneself.

In his chapter, Paul Bernier's motivation is to reconcile, to some extent, Simulation Theory and Theory Theory, and argue that they are about different aspects of our capacity to understand other minds. Bernier sympathises with Gordon's radical simulationism, according to which simulation gives us a direct, nonconceptual entry into the cognitive perspective of others. Gordon's account embodies a non-recognitional model of self-ascription, according to which we do not need any kind of evidence to identify the mental states we are in when we simulate someone. Simply, we use "ascent routines" to answer the question of whether we believe that *p* (a question about the mind) by answering the simpler question of whether *p* (a question about the world as it appears to us).

Now one problem about ascent routines is that they do not seem to be applicable to mental states other than beliefs. However, Bernier points out that a key feature of these routines is that they link the ascription of a mental state with its verbal *expression*. For instance, the child learns to move from the assertion "*p*" to the assertion "I believe that *p*"; she learns how to express her beliefs. Bernier then suggests a way of extending this account to desires.

Radical simulationism faces another, more serious problem. Gordon calls it the problem of "genuine comprehending self-ascriptions". The problem is that we often *reflect on* mental states that we do not express ourselves, for instance when we report or describe them. Gordon himself does not say much about how these reflexive capacities, and the full grasp of mental concepts involved in them, can emerge from practical simulation. Bernier contends that functionalism enters the picture precisely at this point. Simulation must be supplemented by some other cognitive mechanism, of a kind that helps us in developing folk psychology. So radical simulationism yields only part of the truth about mentalisation.

In his reply, Gianfranco Soldati wonders, among other things, whether Bernier's reconciliation between Simulation Theory and Theory Theory leaves us with anything more than a very weak form of Simulation Theory. According to Soldati, Bernier has at best shown how the capacity to make judgements with mental concepts hooks up with practical simulation. What he has not shown is that practical simulation is *necessary* for making these comprehending judgements. So the extent to which our capacity to obtain a direct entry into other minds is relevant to our understanding of mental phenomena is left in the dark.

An important idea that has been only superficially explored in the literature on Simulation Theory is that the first avenue to understanding others is to interpret their actions as having certain identifiable goals. Now in this area of research, an intriguing hypothesis has been put forward in neurophys-

iology: a basic way of understanding an action performed by another subject consists in part in simulating it mentally. Jean Decety's chapter provides a review of the neurophysiological data compatible with such a claim. Perceiving an action seems indeed to be mediated by neural processes also involved in motor executive processes. The way in which the latter are activated depends on the perceptual task. When the perception has no further goal, the pattern of brain activation depends on the nature of the movements presented. But when a subject perceives with the further goal of imitating what she perceives, regions involved in planning and generating actions are activated. Given such an overlap between action-related perceptual and executive systems, the question of motor imagery takes on new theoretical import. A natural hypothesis is that simulating an action helps to plan, evaluate and execute a possible action. Is the perceptual imagery available to the subject imagining an action coded in a way immediately usable in potential actions? This is what is suggested by brain-imagery evidence. Jean Decety reviews the facts (which his own work has helped to establish) suggesting that motor imagery and motor behaviour largely share their neural mechanisms. Mentally simulating an action and actually performing it have the same kind of temporal pattern, adjustment to target width, laterality properties, associated involuntary autonomic responses, and neurological impairments.

Christopher Hoerl's commentary focuses on imitation. To imitate someone, I must have a code that is common to perception and action. The reason is that I must understand what is done in a causal framework, telling me how the observed action is to be replicated. Similarly, imagining performing an action presupposes that my imaginative exercise is causally related to my ability to actually perform that action, and that I am aware of that fact. Hoerl argues convincingly that the phenomenology of imagining an action allows the product to count as objective knowledge.

Pierre Jacob starts with a discussion of the process of mental simulation in a domain to which, as he says, it primarily belongs, namely motor action and imagery. He reviews the empirical evidence for the claim that the ability to perform an action and the ability to simulate it are interdependent. However, Jacob is less enthusiastic about the extension of mental simulation to higher cognitive functions such as those concerned by folk psychology. In particular, he raises a number of problems for Gordon's view that a concept of belief can be extracted from mental simulation itself. Gordon claims that to ascribe a belief is to make a pretend-assertion, that is to make an assertion within the scope of a practical simulation. However, there is a categorical difference between a pretend-assertion, which is at best a representation which *concerns* the

ascribee, and a belief ascription, which is a representation explicitly *about* the ascribee. Mental simulation itself cannot bridge the gap between a representation that merely concerns another person (in John Perry's sense, according to which the reference to the person is left unarticulated) and a representation that is about that person. The ability to engage in pretence is not a sufficient condition for ascribing thoughts to someone else, which requires proper meta-representational abilities.

In his reply, Jérôme Dokic takes over Jacob's argument against radical simulationism. He challenges radical simulationism with a dilemma: either simulation results in pretend motor representations, or it results in more sophisticated states such as pretend propositional attitudes. In the former case, simulation cannot ground an understanding of the contribution to behavior of particular beliefs, while the latter case presupposes the possession of mental concepts. Dokic goes on to argue that mental simulation cannot explain by itself our understanding of a given empirical belief as a "bipolar" mental state, namely something that, in many relevant cases, *can* be true and *can* be false. While Dokic agrees with Jacob that the ability to simulate is not a sufficient condition for being able to ascribe beliefs to someone, he expresses doubts about his stronger claim this is not a necessary condition either.

Jacqueline Nadel's chapter studies the linkages between early imitation and theory of mind. Neo-natal imitation has been presented as expressing an implicit recognition of "like me" actions (Meltzoff & Gopnik 1993). Such a recognition would depend on an innate coupling mechanism, providing a connection between the self and the other, and offering the kind of evidence needed for theory-of-mind acquisition. Nadel examines two informational dimensions that might overlap in Meltzoff's "implicit-recognition" theory: social contingency and imitation recognition. Using a still-face paradigm,[14] Nadel shows that 2-month-olds are able to detect social contingency at the fine-grained level of social synchrony. She shows further that, in contrast with normal 30-month-olds, children with autism of the same mental age do not present any emotional reaction to still-face behaviour in strangers. They do react emotionally, however, if they have prior imitative interaction with the stranger. She thus suggests that implicit understanding of intentional contingency, in autistic as in normal children, might first be restricted to those partners who have already displayed contingent behaviour, then later be generalized to all human partners.

Nadel further brings new evidence to bear on the "like me" feeling attributed to new-borns. Attribution of agency to an imitator is not documented before 9 months; which suggests an asymmetry between "seeing and doing" and "doing to be seen". 30 month-olds tend to use imitation as a commu-

nication device allowing them to express co-reference and topic sharing in a prelinguistic way. That kind of social coupling through object-centred imitation might allow building a repertory of intentional actions to be attributed to self and others. Acquiring such a repertory might be a major step in theory-of-mind acquisition.

Johannes Roessler's comments focus on one of the major claims defended by Jacqueline Nadel, namely the fact that early imitation should not be understood as involving the notion of another agent, even in the implicit "like me" mode. Imitation for Nadel is recognised (rather than only produced) as part of the development of imitation as a communicative device. Roessler discusses the respective roles of communication, imitation and joint attention in the process of theory of mind acquisition. Roessler grants that the practical grasp of other minds allowing a child to use the information presented in behaviour does not entail, prior to communication, grasping psychological concepts or understanding that there are other minds. While granting that imitating others consists in sharing intentions, he defends the view that being aware of the point of the other's action is a case of joint attention. This early ability would allow 12-month-olds to acquire the sense of others as subjects of a point of view. This would prepare them for the next step – taking others to have their own first-person perspective.

François Recanati distinguishes between three forms of simulation. First, simulation is a method for exploiting meta-representations. Consider the process of simulative belief ascription. We run a simulation of someone's beliefs, and reason about the world from her point of view. Such "simulative reasoning" involves two complementary processes: projection and retrojection. Projection allows us to form pretend-beliefs corresponding to the ascribee's own beliefs. Retrojection allows us to move back from the ascribee's world to our own, and explicitly ascribe to her the newly obtained beliefs. In the first part of his paper, Recanati describes some constraints on the correlative processes of projection and retrojection.

Second, Recanati claims that meta-representations themselves are intrinsically simulative. Not only do we use simulation to move from a given meta-representation to another, but a meta-representation is a representation that both refers to a representation and *displays* its content. Meta-representations have an iconic dimension that can be explained in terms of simulation. To entertain a meta-representation, one must be able to simulate the other's perspective, which suggests that one cannot correctly ascribe contents to a radically alien creature. In the context of his discussion of the structure of meta-representation, Recanati makes an intriguing comparison between sentences

like "In Chicago, it's raining" and "According to Pierre, it's raining", conceiving of meta-representation as a special case of a general mechanism that enables us to entertain representations decoupled from the egocentric situation.

Meta-representations are "transparent" in the sense that they are about whichever states of affairs the object representations are themselves about. Recanati then turns to problems raised by situations in which the transparency of meta-representations does not seem to be guaranteed, for instance when there are conceptual divergences between ascriber and ascribee. To deal with these problems, Recanati describes a third form of simulation, which is specifically communicational. This last form of simulation involves "deference" and "context-shifting"; for instance, we move from the actual context to an alternative one in which words are used with a deviant meaning.

Recanati argues that hypothetical reasoning is like simulative reasoning: one takes on, or simulates, contents that one does not believe oneself. In contrast, Jérôme Pelletier doubts that assumption requires simulation, at least if simulation is conceived as a form of imagination. Imagination is participative or perspectival, whereas assumption is a much more detached mental activity. He is also sceptical about Recanati's claim that reasoning on someone else's belief is a kind of simulative reasoning, arguing that such a claim might conflict with the logic of belief (more precisely with the fact that belief is not closed under deduction). Turning to Recanati's thesis that metarepresentation involves simulation, Pelletier suggests that there is no sharp divide between metarepresentations that display the 'object representation' (or primary representation) and metarepresentations that do not. He also thinks that this thesis has difficulties in accounting for the ascription of impossible propositions. Finally, he objects to the claim that metarepresentations are a special case of a decoupling procedure used in other circumstances. Sometimes, decoupling from the actual situation, like in "In Chicago, it's raining", might require imagination in some sense (more precisely, "impersonal imagining"), but there is no reason to think that one has to put oneself in anybody's shoes to understand this sentence.

Donald Peterson discusses the kind of processing style involved in mental simulation. He contrasts monological processing, where only one schema (composed of rules, data and goals) is examined at a time, with dialogical processing, in which two distinct schemas can be handled. He contends that autistic patients have a bias towards monological processing, which may be compensated by developing specialised compound schemas containing exception-clauses and extra detail. Context insensitivity found in autistic persons would therefore affect not only mental tasks, but also non-mental prediction tasks, such as "If Mummy had not moved the chocolate, where would it be?" Adap-

tive modelling of another cognitive system relies, in Peterson's view, on a dialogical strategy of reasoning, combining information on a matrix (the facts as known) and a context. Those children with autism who pass the false-belief task might be taken to have used a monological model including all the relevant (theory-based) psychological inferences, rather than construing through mental simulation a dialogical model for the individual case.

In his reply, Pierre Livet welcomes Peterson's emphasis on the fact that the false belief task requires a specific type of reasoning that might constitute the core difficulty for the three-year-olds and the autistic children who fail it. Livet questions, however, the nature of the problem. He suggests that the distinction relevant for characterising this difficulty is one involving two types of revision rather than two types of processing style. Revising involves changing some of one's beliefs when new information is known, with no additional change being supposed to occur in the world. Updating is the kind of revision needed when the subject is uncertain whether the world did or did not change. In such a case, a subject's reasoning must follow several competing hypotheses, allowing for the possibility that the world may have changed. Maxi's task is for Livet an updating-task. In this picture, what children under four fail to achieve is not so much revision as updating; they tend to accommodate situations that require updating in a simple revision framework.

Joëlle Proust's chapter reviews the pros and cons of Gordon's radical simulationism. Such a view is made appealing by at least three features. First, simulation, in this approach, is performed on the basis of what a subject knows about kinds of context, and only presupposes an ability to evaluate them. Second, simulation is taken to be a wide-ranging psychological process that does not seem to necessarily involve mental concept possession. Third, simulation is harnessed to a practical reasoning system, which allows explanation of converging evaluations across agents.

The difficulty of the view however consists in bridging the gap between this practical ability to predict others and mental concept acquisition and use. A classical objection indeed claims that radical simulationism falls short of explaining independently of Theory Theory how mental concepts are acquired. Proust argues that, although Simulation Theory may not be sufficient for psychological concept acquisition, it is a necessary precondition for it. The specific way in which simulation is involved is that it captures, first, how the world looks different from another subject's standpoint, but also how to use this difference in standpoints to make predictions in the world as it is. Simulating in this sense, according to Proust, cannot be confined to a dynamic process for imaging contexts. A reasoning structure is needed in order to infer conse-

quences within several simulated contexts (at least two) and to examine those consequences that hold *across* contexts.

The idea is that predicting the behaviour of a misinformed agent essentially involves the ability of the simulator to put such behaviour to use in the real world. If Paul lies to Pierre, for example, Paul misrepresents facts to Pierre, probably in order to take advantage (in the real world) of the misleading inferences that will be drawn by Pierre and motivate the latter's actions. If simulation engages such a two-store reasoning capacity, i.e. an ability rather than a piece of theory, the question again is how to articulate it with mental concept possession. Proust suggests that theory and exercise might be two complementary sources in mentalization. The child on the one hand is exposed to a theory that permeates linguistic exchanges between members of her community. She may thus acquire a partial grasp of theory of mind in the shallow way described in ascent routines. On the other hand, in order to fully identify the inferential role of *belief* and other propositional attitudes, she needs to exercise practical reasoning for herself, i.e. deploy in simulation distinct ways of representing situations, and draw inferences within and across contexts.

In his reply to Proust, Pascal Ludwig sides with radical simulationism in denying that self-knowledge is needed to simulate another's situation. Ludwig examines the semantical problems associated with shifting the reference of 'I' when imagining being someone else and putting oneself in her shoes. This procedure of reference shifting seems in some sense acceptable, if one parallels it with the fictional use of proper names (taken as if they had a denotation). This kind of simulation might depend on the general ability described by Vendler of imagining oneself doing things in a subjective, perspectival way. Ludwig does not believe, however, that radical simulationism may provide more than a necessary condition for mental understanding of others. He is sceptical about the possibility to construe mental simulation as emerging from some counterfactual reasoning of the kind described by Proust. Ludwig's own suggestion is that nobody could predict peers' courses of action without knowing at least tacitly the principles of a theory of mind.

John Campbell raises the question of how we understand demonstratives such as "that book" or "that bird" when they are used by other people to refer to some element in their perceptual field. How do I know the thought you are expressing by "I'm going to climb that mountain", referring to a perceived mountain? At the very least, I have to know the causal role of your thought: what prompts you to think it and what consequences you draw from it. Campbell says that the Theory Theory and the Simulation Theory can be viewed as alternative conceptions of the way in which I come to know the causal roles of

thoughts. Do I represent the causal structure of your thought or do I instantiate it myself in the context of simulation?

However, there is a further dimension in my knowledge of what thought you express with a perceptual demonstrative, about which both Theory Theory and Simulation Theory remain silent. This is the thought's normative dimension. To know which is the *right* causal role of the thought, it is necessary to know the *reference* of the perceptual demonstrative. Campbell then claims that a proper understanding of another's use of a perceptual demonstrative requires joint attention. It is not enough for you and me to both attend to the same object; we have to causally coordinate our attention. Joint attention in understanding depends not just on monitoring one's attention but on the common knowledge that we are thinking about the same thing. Campbell argues that although knowledge of causal role need not be conscious, it is crucial that attention itself be conscious.

In her reply, Elisabeth Pacherie suggests that we should distinguish between two different normative issues. The first normative issue is the alignment of the thought's causal role with its reference, and the second concerns the conditions under which mutual knowledge is achieved in communication. Pacherie argues that whereas it is possible to know the reference of another's use of a perceptual demonstrative without engaging in joint attention, the problem of mutual knowledge is more general, and does not concern demonstrative thoughts in particular.

Acknowledgements

We thank heartily the *Ile-de-France Network for Cognitive Science* (RISC) for support on a workshop that resulted in the present book. Also, we want to express our special thanks to Laurence Helleu for compiling the index.

Notes

1. Wimmer and Perner (1983), Wellman (1990), Perner (1991), Gopnik and Wellman (1995).

2. Leslie (1987), Baron-Cohen (1995).

3. See Child (1994).

4. Gordon (1986), Heal (1986), Goldman (1989).

5. Kahneman and Tversky (1982), Goldman (1989), Hutchison, Decety, et al. (1989), (1997), Currie and Ravenscroft (1997), Gallese and Goldman (1998).

6. See Campbell (1994).

7. See Greenwood (1999).

8. Rizzolatti et al. (1988), Hutchison et al. (1999).

9. See Fuller (1995) and Heal (1995).

10. See Perner (1994), Perner (1996).

11. See Leslie and German (1995), Russell (1997), Perner and Lang (1999).

12. Gordon (1995).

13. Rizzolatti et al. (1988).

14. Adapted from Murray and Trevarthen (1985).

References

Baron-Cohen, S. (1995). *Mindblindness, an Essay on Autism and Theory of Mind*. Cambridge: MIT Press.

Child, W. (1994). *Causality, Interpretation and the Mind*. Oxford: Clarendon Press.

Currie, G. & Ravenscroft, I. (1997). Mental Simulation and Motor Imagery. *Philosophy of Science, 64*, 161–180.

Davies, M. & Stone, T. (1995). *Mental Simulation*. Oxford: Blackwell.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). Brain activity during observation of action. Influence of action content and subject's strategy. *Brain, 120*, 1763–1777.

Decety, J., Jeannerod, M., & Prablanc, C. (1989). The timing of mentally represented actions. *Behavioral Brain Research, 34*, 35–42.

Fuller, G. (1995). Simulation and Psychological Concepts. In Davies & Stone (pp. 19–32).

Gallese, V. & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind reading. *Trends in Neuroscience, 2*, 12, 493–501.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language, 4*, 161–85.

Gopnik, A. & Wellman, H. M. (1995). Why the child's theory of mind really is a theory. In M. Davies & T. Stone (Eds.), *Folk Psychology*. Oxford: Blackwell.

Gordon, R. M. (1986). Folk Psychology as simulation. *Mind and Language, 1*, 158–171.

Gordon, R. M. (1987). *The Structure of Emotions*. Cambridge: Cambridge University Press.

Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In Davies & Stone (pp. 53–67).

Gordon, R. M. (1996). 'Radical' simulationism. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 11–21). Cambridge: Cambridge University Press.

Greenwood, J. D. (1999). Simulation, Theory Theory and Cognitive Penetration: No 'Instance of a Fingerpost'. *Mind and Language, 14,* 1, 32–56.

Heal, J. (1995). How to think about thinking. In Davies & Stone (pp. 33–52).

Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, Mind and Logic* (pp. 135–150). Cambridge: Cambridge University Press.

Hutchison, W. D., Davis, K. D., Lozano, A. M., Tasker, R. R., & Dostrovsky, J. O. (1999). Pain related neurons in the human cingulate cortex. *Nature Neuroscience, 2*, 5, 403–405.

Leslie, A. M. (1987). Pretence and representation, The origins of 'theory of mind'. *Psychological Review, 94*, 412–426.

Leslie, A. M. & German, T. P. (1995). Knowledge and ability in "Theory of mind": One-eyed overview of a debate. In Davies & Stone (pp. 123–150).

Murray, L. & Trevarthen, C. (1985). Emotional regulation of interaction between two-month-olds and their mother's. In T. M. Field & N. A. Fox (Eds.), *Social Perception in Infants* (pp. 177–197). Norwood, NJ: Ablex.

Nichols, S. & Stich, S. (to appear). *Mindreading*. Oxford: Oxford University Press

Perner, J. & Lang, B. (1999). Theory of mind and executive function: is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism and developmental cognitive neuroscience*. Oxford: Oxford University Press.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge: MIT Press.

Perner, J. (1994). The necessity and impossibility of simulation. *Proceedings of the British Academy*, 29–44.

Perner, J. (1996). Arguments for a simulation-theory mix. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 90–104). Cambridge: Cambridge University Press.

Rizzolati, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of area 6 in the macaque monkey, II, Area F5 and the control of distal movements. *Experimental Brain Research, 71*, 491–507.

Russell, J. (1997). How executive disorders can bring about an inadequate theory of mind. In J. Russell (Ed.), *Autism as an Executive Disorder*. Oxford: Oxford University Press.

Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge: MIT Press.

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.

Chapter 1

Simulation theory and mental concepts

Alvin I. Goldman
Rutgers University

1. Folk psychology and the TT-ST debate

The study of folk psychology, on the approach I favor, is the descriptive epistemology of folk mentalization. By "descriptive epistemology" I mean an inquiry into the mechanisms and processes of folk mentalization, whatever they turn out to be.[1] By "folk mentalization", I mean ordinary people's attributions of mental states (especially the propositional attitudes) to both self and others. "Descriptive" epistemology contrasts with "normative" epistemology. In this context, the question is not what *justifies* the folk in making their mentalistic attributions; we simply ask what is involved in the generation of these attributions. By calling the study of folk psychology a branch of epistemology, I mean to alert readers to the fact that the questions on the table are not metaphysical ones. We do not attempt to determine the actual constitution of desire states, belief states, and so forth. We are only interested in how the folk understand or conceive of these states, and how they go about attributing them. The job will be done correctly if we get the story of folk mentalization right, even if it turns out that there are no such things "in reality" as desires and beliefs, or even if it turns out that the folk are importantly misguided or underinformed about them.[2] The job might be done correctly even if it turns out that there are other kinds of evidence for mental states (e.g., scientific evidence) that the folk do not use at all or do not use correctly.[3]

The following questions comprise the chief items on the agenda for the study of folk psychology:

1. How do adults execute the tasks of making (A) third-person attributions of mental states and (B) first-person attributions of mental states?
2. How is their competence or facility at making these attributions acquired?

3. What are the concepts of mental states that children and adults deploy? That is, what do they understand mental states to be? How do they conceptualize them? Since propositional-attitude states involve both attitude-types and attitude-contents, this question decomposes into questions about attitude types and questions about attitude contents.

Answers to some or all of these questions have been proposed by both philosophers and cognitive scientists. A full inquiry therefore requires consideration of findings and literatures in both fields. Thus, the study of folk psychology should be a joint venture between philosophers and cognitive scientists.

In recent years a principal debate has taken place between two approaches to folk psychology: theory theory and simulation theory. The debate has focused on question (1A) listed above, the question of third-person mental-state attribution. But evidence has often been invoked that is more directly relevant to other questions on the list. So although the TT/ST controversy centers on question (1A), the controversy broadens out considerably from there. The situation is a bit untidy, moreover, because TT has a rather unified set of answers to all questions on the list, whereas ST mainly addresses question (1A).[4] To get a full-scale debate, ST must be supplemented by answers to the other questions, and simulationists do not fully agree about those answers. These remarks are not intended to favor TT. Although TT can offer unified answers to all the questions, I do not find those answers very plausible. Furthermore, although a loose unity for TT can be descried, that unity hides great diversity in details, and the unity is difficult to maintain. Perhaps it is well to begin with the general contours of TT and some of the problems it faces.

2. Theory Theory and its problems

Generally speaking, theory-theorists maintain that the folk's concepts of the attitude types (belief, desire, and so forth) are theoretical concepts. By a theoretical concept I mean, roughly, the concept of a state defined in terms of causal-inferential relationships to publicly observable (perceivable) events, in this case the environment and overt behavior of the agent. This construal of the attitude concepts is particularly clear for analytical functionalism and conceptual role semantics.[5] Admittedly, philosophical functionalism is typically silent about question (1) on our list: how are particular interpretations actually executed? But some interpretation theorists – e.g., Dennett with his "in-

tentional stance" – certainly suggest that interpretation takes place by means of theoretical reasoning guided by the posited interpretational constraints.

In developmental psychology, theory-theorists tend to be functionalists, although they rarely use that term. They often claim explicitly that mental attribution is executed by means of theoretical inference, in both third-person and even first-person cases. (Gopnik 1993 is especially clear on this point.) Their chief evidence for the theoretical-inference position, however, stems from their answers to questions (2) and (3) on our list. Developmentalists find evidence of children's performance changes in mental-state attributions, and these performance changes are widely interpreted as changes in conceptual repertoires. From these supposed changes in conceptual repertoires, it is inferred that the concepts in question must be theoretical ones.

Since this is the main empirical evidence for the TT position, and since new evidence has emerged in recent years, let us look at it more closely.[6] The orthodox position is that children move from a nonrepresentational understanding of mind to a representational understanding. This orthodoxy, however, has two variants. On one view, children's grasp of the very same mentalistic term changes over time. For example, children initially understand 'belief' in a nonrepresentational fashion; but later (e.g., by age four), they understand 'belief' representationally (Gopnik 1993). On a second view, there may be no single mentalistic term that is conceptualized differently at different times. Rather, at certain early ages only nonrepresentational concepts are acquired, and representational concepts are mastered later. Thus, Bartsch and Wellman (1995), who have amassed the most extensive data on children's talk about the mind, deny that belief is first grasped nonrepresentationally and later representationally.[7] Nonetheless, Bartsch and Wellman retain allegiance to a basic progression from nonrepresentational to representational concepts. They claim that before age three only nonrepresentational concepts are grasped (e.g., the "simple", nonrepresentational concept of desire), and the representational concept of belief is acquired thereafter. Similarly, Perner (1991) appears to hold that no truly representational concept is acquired before age four, and that until that time children do not really understand the concept (any concept) of belief. Since there are these two variants of the orthodox TT position, I shall examine them separately and identify difficulties for each.

The first view is supposedly supported by changes in children's competence at false-belief tasks. It is assumed that this change in competence can be attributed to a change in the conceptual grasp of belief, from a nonrepresentational to a representational understanding. The tenability of this assumption, however, has been undercut by accumulating evidence that young children's

problem with false-belief tasks is not a *conceptual* deficit. A number of studies suggest that 3-year-olds' fragile memories for narratives are central to their poor performance on the traditional false-belief task. Three-year-olds succeed on these tasks when they receive an aid to memory, or when they go through a story twice (Mitchell & Lacohee 1991; Lewis et al. 1994; Freeman & Lacohee 1995). Russell's experiments on the "windows task" suggest that the failure of 3-year-olds on false-belief tasks may result from an "executive" deficit rather than a conceptual one, where an executive deficit consists in an inability to inhibit a prepotent but wrong action (see Russell et al. 1991; Russell 1996: 132–138). Three-year-olds may be incapable of inhibiting a report of a true fact, which is more salient to them, though they understand that the target child has a false belief contrary to that fact. In a similar vein, Clements and Perner (1994) found evidence of an implicit grasp of false belief even when explicit answers to false-belief questions do not reveal understanding. Finally, as indicated earlier, Bartsch and Wellman's (1995) extensive data fail to provide evidence that 3-year-olds have a merely nonrepresentational construal of belief; in other words, it fails to support the thesis that 3-years-olds cannot conceptualize false belief.

Let us turn, then, to the second variant of the developmentalists' TT position, which depends critically on the contention that the early-acquired concept of desire is a purely nonrepresentational concept. Flavell (1988), Perner (1991), and Wellman in several works contend that 2-year-olds have only a simple ("connectionist") grasp of wanting, involving a relation between the target person's mind and some object or state-of-affairs *in the world*. Wanting is not construed by 2-year-olds as involving a mental representation of objects or states of affairs. As Wellman puts it, "the conceiver can simply think of the target person as having an internal longing for an external object" (1990: 211).

By my lights, there are two serious problems with this approach. First, it is universally agreed that understanding false belief is a sufficient indicator of a representational concept of belief. Unfulfilled desire is precisely analogous to false belief; the only difference is the "direction of fit" between mind and world. Since Bartsch and Wellman clearly admit that 2-year-olds understand the possibility of unfulfilled desires, the parallel conclusion to draw is that 2-year-olds have a representational concept of desire. True, desires do not represent in exactly the same way that beliefs do. Desires do not depict the way the world *is*, but how the world *should be* (as judged by the agent's preference). But this simply is the direction-of-fit difference between desire and belief. It is not a reason to deny representational status to a desire concept. Second, Bartsch and Wellman do not adequately solve the problem of the supposed relata of unfulfilled desires, according to the 2-year-olds' putative nonrepresentational model.

They admit that among the "objects" of desire that 2-year-olds conceive of are states-of-affairs that go unfulfilled. But they do not explain what actual-world objects these unfulfilled states-of-affairs might be. So the evidence in support of a distinctive stage of mentalistic conceptualization is very thin indeed. At this point, the case for conceptual change (as opposed to performance change) in childhood mentalisation is actually rather weak. Moreover, even if conceptual change were demonstrated, it would take further argument to establish that this conceptual change reflects a change in *theory*.

I have been discussing problems with the supposed evidence in support of a TT answer to question (3), at least in terms of children's mentalization. Now let me turn to a different problem with the TT position: its stance on first-person attribution (question (1B)).[8] The notion that my self-attribution of current thoughts, intentions, and other mental processes stems from theory-mediated reasoning has little surface plausibility. I now attribute to myself an intention to snap my fingers in the next 10 seconds. What theoretical inference, based on what observed data, could lead me to this self-attribution? There are no observed conditions in my environment that initiate this inference; nor is there any observed overt behavior of mine that prompts it. Similarly, at this moment I am retrieving from memory the year in which I first gave a paper in Paris, and I attribute to myself this bit of memory retrieval. What is the currently perceived evidence from which I might make a theoretically guided inference to the occurrence of such a memory event? Clearly, there isn't any; so an entirely different story needs to be told.

Aware of the serious problems facing TT in this arena, Nichols and Stich (forthcoming) have recently modified their heretofore unqualified endorsement of the TT approach. They now propose a "direct", non-theory-mediated account of self-awareness, which they call *monitoring*. Although this is certainly a step in the right direction, there is a major gap in their overall approach, symptomatic of a central problem facing TT. This is the problem of reconciling a "direct detection" account of self-awareness and a theory-based account of the concepts of the attitudes. On a TT account of attitude concepts – for example, a functionalist account – it is quite unclear how a person can directly detect a state's being an instance of a belief or a desire. Being a belief or being a desire consists in having an appropriate set of dispositions to interact with other such states and with environmental inputs and/or behavioral outputs. How is an instantiation of such relational/dispositional properties supposed to be directly detected (or "monitored")? I posed this problem in Goldman (1993), and it has not yet been satisfactorily answered by defenders of TT. Since functionalism says that any particular attitude state is type-identified by its (actual and

possible) relations to other such states, and those other states are in turn type-identified by their (actual and possible) relations to further such states, there is a serious threat of combinatorial explosion. This problem is camouflaged by Nichols and Stich's discussion because they resort to the oft-used metaphor of beliefs and desires being occupants of cognitive "boxes". Talk of boxes, they acknowledge, is simply shorthand for a functionalist story. But if functional properties are what type-identify particular mental states, it is radically unclear how simple "monitoring" can determine the requisite types.[9]

3. Simulation theory

The last point identifies one of the main issues I shall be taking up in this paper. Before returning to that issue, however, I want to explain the principal alternative to TT vis-a-vis question (1A): the simulation theory.

TT portrays a folk attributor as reaching an attributive conclusion (e.g., "Black will decide to castle on his next move") by purely factual reasoning. Factual reasoning is here construed as a sequence of causally related steps each of which consists in a belief (or other doxastic state). This factual reasoning is guided, or aided, according to TT, by belief in some sort of principles of human psychology. These principles might consist of (believed) psychological laws; if charity theories are construed as a species of TT, the principles might consist of principles of rationality, such as "People generally maintain coherent systems of belief," or "People generally desire what it is good for them to have". At a minimum, TT does not postulate any non-doxastic types of states in the attributor's own procedure of determining the target's states.

ST departs fundamentally from TT in that it does posit non-doxastic states as steps in the attributor's typical procedure. The basic idea of ST, of course, is that the attributor tries to mimic the mental activity of the target. In the special case of mimicking a target's factual reasoning, this will result in only doxastic states (including pretend belief states) occurring in the attributor. In all other cases, however, non-doxastic mental states may be used essentially by the attributor. In arriving at an attribution of a decision, for example, the attributor will herself deploy (pretend) desires in trying to mimic what the target will decide. This will lead to a (pretend) decision in the attributor, another non-doxastic state. Finally, when this (pretend) decision is reached, the attributor uses that decision to form a belief about the target's decision.

The second point of contrast between ST and TT, already implicated in the foregoing, is that ST posits intensive use of *pretend* states, which TT has no use

for at all. By a pretend state, I mean some sort of surrogate state, which is deliberately adopted for the sake of the attributional task, and does not ultimately play the same causal role in the attributor's psychology that states of that sort normally play. It is a basic premise of ST, however, as I construe it, that pretend states have many of the same causal powers as ordinary, non-pretend states with the same contents. This property is essential if simulation is to be capable of successful mimicry. If pretend desires do not have at least roughly similar effects in practical reasoning as genuine desires, simulated practical reasoning won't generate accurate decisional attributions, even if the right pretend desires are used as inputs. Of course, ST is not committed to any strong claims of accuracy for the simulation heuristic (which can go awry in various ways). But it must account for at least a modest level of accuracy, since that is what interpersonal mentalization appears to achieve.

One further feature must be added to ST if it is not to "collapse" into a form of TT. If an attributor's procedure of simulation is guided by theoretical beliefs about the target's psychology, it would seem to reduce to a variant of theoretical reasoning. (It still wouldn't be a standard form of theoretical reasoning, however, since some steps in the procedure would be non-doxastic states.) The standard form of ST postulates that simulation is not driven by theoretical principles; rather, as I have phrased the matter, it is "process-driven" (Goldman 1989). In simulating practical reasoning, the attributor feeds pretend desires and beliefs into her own practical reasoning system and lets that system do the work of churning out a choice or decision. This frees the attributor from the necessity of relying on theoretical principles of psychology.[10]

Having drawn a reasonably sharp contrast between ST and TT, we must now acknowledge that there is not just a *single* version of ST but *many* possible versions that vary in strength. One dimension of variability is *frequency* of use. The strongest form of ST would say that all cases of (third-person) mentalization employ simulation. A moderate version would say, for example, that simulation is the *default* method of mentalization but can be supplemented or overridden by theoretical considerations. A weak version would say that the simulation heuristic is used only occasionally. I am attracted by the moderate version. Certainly *some* cases of mentalization do not utilize simulation. If I know from previous cases that Thelma always chooses raspberry when offered ices or sorbet, I predict that she will make that choice again on the present occasion. I have no need for simulation. This inductive (if not clearly theoretical) method of mentalisation is one I have always granted (e.g., Goldman 1989). If this makes me a hybrid theorist, so be it.

Another dimension of variability is the *source* of simulation. It might be thought of as a mere short-cut, or what Fuller (1995: 21) calls an "epistemological tool". Perhaps we conceptualize the attitudes in theoretical – e.g., functionalist – terms, and therein accept certain functional laws. But we learn through experience that simulation is a good short-cut to theoretical reasoning with functional laws. This is a possible version of ST (especially if simulation is process-driven); but it is obviously weaker than a version which holds that (1) simulation is the primitive, root form of interpersonal mentalization, and (2) mental concepts are not purely theoretical concepts. I tend to favor this latter, stronger version.

A third dimension of variability is the *scope* of simulational mentalisation. Heal (1996), for example, wishes to restrict ST to the claim that simulation is used for attributing contentful states based on rational relations. She writes:

> The kind of simulationism I would like to defend says that the only cases that a simulationist should confidently claim are those where (a) the starting point is an item or collection of items with content, (b) the outcome is a further item with content, and (c) the latter content is rationally or intelligibly linked to that of the earlier item(s). (1996: 56)

Other simulationists might wish to claim that simulation is also (successfully) used to attribute emotions and sensation states (e.g., feeling tense, nervous, or upset), where there is no "rational" relation between the input and output states.[11] In other words, simulation might proceed by feeding pretend inputs not only into factual and practical reasoning mechanisms but also into a wider class of mental-state generating mechanisms.

4. Empirical plausibility of simulation theory

Since theories of folk psychology, as I contended earlier, are appropriately viewed as empirical theories assessable by cognitive science, we must inquire into the empirical plausibility of ST. At this juncture I am not looking for clear-cut evidence that favors (some form of) ST over (competing forms of) TT, but merely the viability of ST as judged by cognitive science. The comparative popularity of its competitor, TT, doubtless stems in part from the fact that, as Stich and Nichols (1992) put it, the "dominant explanatory strategy" in cognitive science is to view cognition as the acquisition and application of internally represented knowledge structures. These same authors regard ST as potentially "an insidious threat to traditional theories in cognitive science" (Nichols et al.

1996: 40). However, this notion that simulation would be an island in a sea of theorizing is entirely misplaced. At any rate, simulation is an intensively deployed type of heuristic in many sectors of human cognition. When this fact is appreciated, its plausibility should have a better ring in the ears of cognitive scientists.[12]

Mimicry or synchrony is a widely employed device in Nature's bag of cognitive tricks. It is found in a wide range of physical, physiological, and mental dimensions of human life (for a review, see Levenson & Ruef 1997). Young children and adults revel in mimicry rituals whether they involve postures, vocalizations, or facial expressions. The generic trick I wish to highlight, however, has an epistemic, or information-gathering, theme. I shall call it *investigation through re-creation*. Human beings engage in a variety of behaviors, both overt and purely mental, in which they seek to "re-create" observed or anticipated scenarios. I conjecture that the function of such behaviors is to investigate, study, or explore the actual or possible scenarios that they model. I do not say that these re-creative activities are always executed with a conscious epistemic aim. Rather, we have propensities to engage in such re-creative activities because they have, as a frequent by-product, the disclosure or discovery of useful information. Mental simulation, I conjecture, is one specimen of this generic type of heuristic.

Here is a brief and compressed sampling of what I have in mind. (1) *Direct behavioral imitation*. Meltzoff and Moore (1977, 1983) discovered that newborn infants have an innate ability to imitate facial expressions, and imitation remains an important proclivity among older infants and young children. Clearly, direct behavioral imitation re-creates an observed scenario. (2) *Pretend play*. Pretend play is not direct imitation, but it is quasi-imitation of a genre of activity. When a child pretends to talk on a telephone by using a banana as a prop (Leslie 1987), he is not precisely imitating normal telephone usage. But the (overt) pretense does re-create selected aspects of the modeled activity. (3) *Drama*. Similarly, the entire idiom of drama is a medium in which people imitate or impersonate individuals, typically not actual individuals or events but persons and scenarios that might exist or occur. The apparent cross-cultural universality of the dramatic idiom throughout recorded time strongly suggests an innate basis for this kind of activity.

I turn now to purely mental re-creation. (4) *Visual imagery*. If I visually imagine the rooms in my house, I re-create (to some approximation) the visual perceptions associated with being in those rooms and looking around. Prospective re-creation occurs when one engages in planning activity. A person in the process of climbing a tree or a cliff may imagine how he would be

positioned if, in his next move, he places his left foot onto a particular branch or ledge. Where would that move leave him?

These forms of re-creation are valuable because they characteristically yield useful information. Visually imagining the rooms in one's house can yield information about the layout of the rooms, or how many windows are in the house. Imagining the result of a prospective climbing maneuver can yield information about the effectiveness or ineffectiveness, the riskiness or non-riskiness, of that move.[13] Attempts at behavioral imitation can yield other types of information. It can help the imitator learn what is required, in motor terms, to reproduce the target action. And it can help the actor discover the physical and social consequences associated with that type of behavior, including responses from other members of the actor's community. These are all valuable pieces of information, for children especially. Pretense has similar epistemic payoffs, at least in many cases. If a child practices talking on the telephone in pretend situations, she will be better skilled at doing the real thing and will make fewer mistakes when it "counts". She may learn the expected elements and proprieties that govern and facilitate telephonic conversation. Impersonation and drama help people learn how circumstances and situations different from their own will be experienced.

These kinds of re-creative activities disclose useful information that help creatures chart a course through their physical and social environments. This is why they proved adaptive and why evolution wired us up with these propensities. Gopnik and Meltzoff (1997) make a convincing case for the thesis that infants and children are inquiry-driven creatures, a characteristic that surely has a genetic basis. I am suggesting that children's dispositions to engage in a variety of re-creative activities are specimens of this trait of inquiry-drivenness. The next suggestion, obviously, is that mental simulation might have similar roots. If mental simulation enables us – often if not always – to make correct mental-state attributions to others, this epistemic payoff would be useful enough to have been favored by evolution.

As we noted earlier, however, simulation cannot produce accurate third-person attributions unless pretend states resemble their non-pretend counterparts in crucial respects. In particular, they must generate the same outputs as their counterparts when operated upon by the same pertinent processing mechanisms. For this to hold, they presumably must share cognitive and/or neurological properties. Is this an empirically defensible hypothesis about pretend states and their non-pretend counterparts?

At present there is no evidence concerning pretend and non-pretend beliefs or desires in general, at least to my knowledge. But very probative evidence is

available for other cognitive domains. For example, if visual imagery is understood as pretend vision, then there is excellent evidence for the sharing of neural resources as between pretend and non-pretend vision (Farah 1988). There is additional dramatic evidence in the domain of motor representations. Motor imagery can be construed as pretend motor intentions or plans, in which the representation of a given motor act is internally rehearsed without any overt motor output. Such simulation of action planning has been found to have striking functional equivalences with actual preparations to move, and to share the same neural substrate (Decety & Grezes 1999; Decety & Jeannerod 1996; Decety et al. 1994; Stephan et al. 1995). It was also found that motor imagery activates heart and respiration control mechanisms in proportion to the actual effort that would be required for a real action (Decety et al. 1993). If pretend and non-pretend states in the motor domain can correspond so closely, why shouldn't there be analogous correspondences for propositional attitudes in general?

5. Simulation and an 'internal' grasp of the attitudes

In this final section I return to question (3) on our list, the question of how we understand or grasp the attitude *concepts*. At the end of Section 2, I pinpointed difficulties that arise about first-person attribution if one posits a purely theory-based – especially functionalist – account of the attitude concepts. First-person attribution seems to involve some sort of monitoring, introspection, or direct detection, but it is hard to make such an account work if the attributor understands 'desire' and 'belief' in purely behavioral or functional terms. ST has a further reason to avoid this theory of the attitude concepts, since it would concede a significant chunk of territory to TT.[14] Is there any alternative? A tentative proposal I have floated (Goldman 1989, 1993) is that the attitudes might be grasped in phenomenological terms. The idea is that occurrent attitude states have phenomenologies, and different attitude types are conceptualized in terms of their distinctive phenomenologies.[15] If one is unpersuaded by such a phenomenological proposal, a second sort of "internalist" account is available. Perhaps there is a class of intrinsic (non-relational) features at the sub-personal level that distinguish among the attitudes. Perhaps these features are what get associated with words like 'want' and 'think'. If such features are directly detected and hence (internally) "observed", this offers a second "internalist" alternative to the standard TT approach.[16]

Unfortunately, major roadblocks seem to confront such an account. Isn't there evidence that very young children detect certain types of attitude states – specifically, desires, goals, or intentions – in the physical behavior or motion of others? At a minimum, infants are specially attuned to human behavior that adults would describe as purposive or goal-directed. This suggests that infants' basic grasp of these attitudes is in *behavioral* terms, not in terms of internal features, whether they be phenomenological or sub-personal. According to Premack (1990) and Gergely et al. (1995), for example, the core notion of intentionality is present innately and is "triggered" by certain perceptual patterns, e.g., self-propelled motion or motion with respect to a potential goal. Baron-Cohen (1995) seconds Premack's idea by positing an "intentionality detector" (ID), which he claims to be an innate endowment for reading mental states in behavior (1995: 32). As described by Baron-Cohen, ID interprets almost anything that propels itself or makes a non-random sound as an agent with goals and desires (1995: 34).[17]

In an admittedly speculative fashion, I shall approach this problem by advancing a *dual-representation* hypothesis about mental state concepts. I shall focus on goals, desires, or intentions, though I suspect that many of these proposals apply equally to other types of mental states. The rough idea is that people develop *two* sorts of mental representations for mental states: representations of behavioral characteristics and representations of internal characteristics (whether the latter are phenomenological or non-phenomenological). In tendering this hypothesis, I do not mean to propose that people have one type of representation for third-person desires and a different type of representation for first-person desires. That would yield the traditional puzzle in philosophy of mind of how they can conceptualize both themselves and others as having the same sorts of states. Rather, they come to understand certain (roughly) *behavioral* representations and certain *inner-feature* representations as representations of one and the same (sort of) state. Another component in my story is that there is indeed change and development in the grasp of mentality at different (early) ages, but the changes I have in mind do not coincide with those standardly discussed in the developmental literature. As noted above, there are signs that young infants (perhaps as early as five months) detect intentionality or goal-directedness in the behavioral patterns of other persons. At that point they may not yet recognize certain internal features associated with their own experiences of desiring and intending as marks of these desires, purposes, or intentions. Under the present hypothesis, however, there is a later stage at which children begin to construe the two types of representations as representations of one and the same (type of) state. That is when they achieve a fairly developed

conception of mental states, at least within hailing distance of the adult conception. Indeed, if their initial conception of intentionality is altogether devoid of the postulated "inner characteristics", one might well say that at that juncture their conception is merely a "proto-concept" or "Ur-concept" of mentality.

The idea of dual, or multiple, representations of a single type of object or category is quite common in cognitive science. For example, people might represent a single sortal in terms of both shape and function; and they may deploy representations utilizing different cognitive codes or modalities. Thus, Biederman (1987) proposes that the classification of visually observed objects exploits representations that code object-types in terms of their shapes. Identifying something (visually) as a piano or a lamp exploits a stored model or prototype of what pianos or lamps look like. In addition, there are likely to be separate, modality-neutral codes that represent pianos and lamps in terms of their functions, e.g., "keyboard musical instrument" and "portable source of (artificial) light". Another example is the postulation of different codes associated with the ventral and dorsal anatomical pathways of vision. The dorsal pathway is thought to employ a code for the guidance of action in relation to seen objects, a code that utilizes visual information relevant to grasping objects, for example (Milner & Goodale 1995).

Some pairs of matching representations are antecedently quite surprising. Meltzoff and Moore (1983) discovered that infants as young as forty-two minutes can imitate the facial gestures of another person. This suggests that newborns not only represent their own facial movements proprioceptively (after all, they do not *see* their own faces), but also they must somehow *match* those proprioceptive representations to entirely different, visual representations of the target's facial gestures (Meltzoff 1999).[18]

I now wish to conjecture that children learn to match inner features detectable in their own goal or desire experiences with behavioral cues utilized in identifying goals or desires. In principle, this could be achieved by noting their own goal-driven behavior and its regular association with these inner features. But this would leave us with the traditional question of why they should think that *other* people ever undergo analogous inner experiences. This would be a particularly vexing problem for ST, because ST postulates that third-person attributors make an inference from their own pretend internal states (the output states of their simulations) to corresponding *inner* states of their targets. But why should they think that other people ever undergo such internal states at all? Philosophers of mind have shied away from internal features in their theorizing precisely because of this looming problem of other minds.

At least a partial answer might be provided by the phenomenon of "mirror neurons".[19] In studies of macaque monkeys, an interesting class of premotor neurons have been discovered that discharge both when the monkey performs an action and when it observes another individual making a similar action (Rizzolatti et al. 1996). The discharge of these neurons is not correlated with individual movements, but rather with general types of actions, especially grasping, holding, tearing, poking, and manipulating. So these neurons seem to code motor representations of goal-directed actions involving objects. Mirror-neuron activity, then, is apparently a neural correlate of plans or goals for action. Surprisingly, mirror neurons are also triggered when a mere observer watches a target agent act toward a goal. This correspondence between observed and executed actions is sometimes described in terms of the metaphor of physical "resonance" (Rizzolatti et al., forthcoming). It is as if neurons in certain motor areas start to "resonate" with their cousins as soon as the appropriate visual input is presented. This resonance phenomenon is not restricted to monkeys, but is found in humans as well, as confirmed by transcranial magnetic stimulation (TMS), MEG/EEG recordings, and brain imaging techniques. Experiments demonstrate that motor centers of adult humans resonate during movement observation. For example, Fadiga et al. (1995) stimulated the left motor cortex of normal subjects using TMS while they were observing arm movements and hand grasping movements performed by the experimenter. Motor evoked potentials were recorded from various arm and hand muscles. A selective increase of motor evoked potentials was found in those muscles that the subjects normally use for producing the observed movements. Findings in clinical populations further support these ideas. Lhermitte et al. (1986) document a phenomenon of "imitation behavior", in which patients with prefrontal lesions compulsively imitate gestures or even complex actions performed in front of them by an experimenter. This behavior is explained as arising from an impairment of inhibitory control that normally governs motor plans. Apparently, when observing someone else perform an action, a normal human generates a "resonant" plan or an image of doing that action himself. This plan is normally inhibited so that it does not yield motor output; but such inhibition is impaired in the relevant patient population. The same phenomenon was called "echopraxia" by Dromard (1905).

Gallese and I (Gallese & Goldman 1998) have cited resonance phenomena as possible evidence for, or a precursor of, mental simulation. But here I mean to invoke them for the purpose sketched above, viz., to proffer a possible explanation of how a child might come to interpret other people as undergoing internal experiences of desiring or planning similar to his own experiences.

Such an interpretation could be facilitated by the occurrence of resonant experiences produced while observing others. After all, the observer can recognize that he does not himself act when he undergoes this experience; and in typical situations there isn't even an appropriate goal object for him to act upon. Furthermore, he sees that the person he is watching *is* acting in the appropriate way. So it might be easy for him to attribute the motoric experience he himself undergoes to the other actor. In other words, resonance activity might psychologically facilitate or promote the attribution to others of experiential (or internal) dimensions of mental states. As we have seen, this is important if simulation is to be a correct account of how third-person mental attribution transpires.[20]

Notes

1. There is an older use of the phrase 'folk psychology' in which it refers specifically to a supposed body of commonly known laws that people use in the process of ascribing mental states. On the present usage, this is only one hypothesis about how people mentalize, one that is very much up for debate. Thus, the present usage is intended to be neutral about the existence and centrality of such putative laws.

2. However, since folk mentalization is an activity of attributing mental states, and attribution itself seems to consist (at least in part) of making judgments or having beliefs, our target task seems to commit us, at least provisionally, to the existence of beliefs.

3. Many writers about the attitudes are more interested in the metaphysical questions than the folk-epistemological ones, even if it looks like they are working on the problem of folk psychology. For example, in his discussion of "radical interpretation," David Lewis writes:

> It should be obvious by now that my problem of radical interpretation is not any real-life task of finding out about Karl's beliefs, desires, and meanings. I am not really asking how we could determine these facts. Rather: How do the facts determine these facts? By what constraints, and to what extent, does the totality of physical facts about Karl determine what he believes, desires, and means? (Lewis 1983: 110)

Clearly, the problem that interests Lewis is the metaphysics of the attitudes, not the question of how the folk form concepts and beliefs about the attitudes. Somewhat similarly, when Jerry Fodor offers us a representational theory of mind, telling us, e.g., that beliefs are relations to sentences in mentalese, he is not making a contribution to folk psychology in my sense but to the metaphysics of the propositional attitudes.

4. Gordon's "ascent-routine" approach to the understanding of the attitudes is an exception to this claim. But not all simulationists would agree with this application of the simulationist idea.

5. It might also be extended to charity theories, which add further constraining relationships such as rational relations among the attitude states and between those states and the agent's circumstances.

6. My discussion is excerpted from Goldman (2001).

7. They write: "[O]ur data provide no evidence that a representational understanding of beliefs is a significantly later achievement, following only on the heels of an earlier "connections" misconstrual of beliefs. Instead, the data suggest that very soon after children begin to talk about thoughts at all, they discuss a variety of distinctions among mental contents and states of the world, including some that seem to presuppose a representational understanding of mental contents as separate from but about the world." (1995: 57)

8. Some theory-theorists, e.g., Gopnik (1993), use evidence about first-person attribution to support the TT position, but I find this evidence unconvincing. There is no space to review my reasons here. See Goldman (2000, 2001).

9. Nichols and Stich (forthcoming) advert to the problem of detecting attitudinal-state contents in the context of criticizing my introspectionist account. I grant that I have not satisfactorily solved this problem, but this should not cloud their own difficulties with the problem that their (largely tacit) functionalism poses.

10. This is not to say that no assumptions lie behind simulation at all. There might, for example, be a tacit belief by the attributor that "my psychology works the same as the target's psychology". I do not think that a theory qualifies as a version of TT if it posits such a "weak" tacit belief on the part of the attributor.

11. For example, if you are trying to determine the emotional state of someone crawling on their stomach through a dark cave, with only a few inches of room above their head, you might imagine yourself in that situation and see what emotional state is triggered in you. (This example is due to Walton 1997.)

12. The discussion in this section is excerpted from Goldman (2001).

13. The value of strategy-testing has been offered by Currie (1995) as an evolutionary explanation of imagination and of (mental) pretense. Currie in turn credits Dennett (1978). Neither of these authors generalizes this sort of explanation, as I am doing, to all varieties of re-creative activity.

14. In addition, third-person simulation seems to need a direct-detection account of the outputs of simulational processes.

15. My discussions in Goldman (1989, 1993) tended to suggest a phenomenological or at least internalist account of attitude contents as well as attitude types. That is how Nichols and Stich (forthcoming) and Peacocke (1998, 1999) interpret me. I really intended to be rather non-committal about the contents. At any rate, the question of attitude contents is set aside for the present paper.

16. Whichever of these options is pursued – the phenomenological or the non-phenomenological option – some story must be included about the features by which the contents of the attitudes are detected. That story might include the "language of thought", for example. However, this topic goes beyond the scope of the present paper.

17. Some aspects of these ideas are challenged by a study of Woodward (forthcoming). Woodward found that 9-month-olds differentiated between movements of grasping a toy and movements of merely letting the back of one's hand drop onto a toy. In other words, by nine months of age infants selectively encode certain goal-related properties and not others; and this selective encoding appears to begin at roughly five months. So infants don't seem to attribute goal-directedness indiscriminately to any motion of a self-propelled entity. However, these sorts of qualifications still leave in place the fundamentally behavioral character of the cues used for third-person goal-state detection.

18. In this case, of course, the infants' "matching" representations are not representations of numerically the same action, only similar actions (their own and that of the observed target).

19. The reader will please remember that I am not addressing the traditional problems of normative epistemology in the mental domain. I am not asking what confers justification on people's beliefs about other minds. I only seek a descriptive account of the generation of their beliefs, which is what the theory of folk psychology, as I construe it, aspires to provide.

20. Thanks to Bill Child for his very helpful commentary on the original version of this paper (which led to several revisions), and to members of CREA who participated in the discussion at the conference.

References

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

Bartsch, K. & Wellman, H. M. (1995). *Children Talk About the Mind*. New York: Oxford University Press.

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Clements, W. & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*, 377–395.

Currie, G. (1995). Imagination and simulation: Aesthetics meets cognitive science. In M. Davies & T. Stone (Eds.), *Mental Simulation*. Oxford: Blackwell.

Curie, G. & Ravenscroft, I. (1997). Mental simulation and motor imagery. *Philosophy of Science, 64*, 161–180.

Decety, J. & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences, 3* (5), 172–178.

Decety, J., Chaminade, T., Grezes, J., & Meltzoff, A. N. (2002). A PET exploration of the neural mechanisms involved in reciprocal imitation. *Neuroimage, 15* (1), 265–272.

Decety, J. & Jeannerod, M. (1996). Mentally simulated movements in virtual reality: does Fitt's law hold in motor imagery? *Behavioral Brain Research, 72*, 127–134.

Decety, J., Perani, D., Jeannerod, M., Bettinardi, V., Tadary, B., Woods, R., Mazziotta, J. C., & Fazio, F. (1994). Mapping motor representation with positron emission tomography. *Nature, 371*, 600–601.

Decety, J., Jeannerod, M., Durozard, D., & Baverel, G. (1993). Central activation of autonomic effectors during mental simulation of motor actions in man. *Journal of Physiology, 461*, 549–563.

Dennett, D. (1978). Why the law of effect won't go away. In *Brainstorms*. Montgomery, VT: Bradford Books.

Dromard, G. (1905). Etude psychologique et clinique sur l'échopraxie. *J. Psychol.* (*Paris*), *2*, 385–403.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Farah, M. (1988). Is visual imagery really visual? Overlooked evidence from neurophysiology. *Psychological Review, 95*, 307–317.

Flavell, J. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In J. Astington, P. Harris, & D. Olson (Eds.), *Developing theories of mind*. New York: Cambridge University Press.

Freeman, N. & Lacohee, H. (1995). Making explicit 3-year-olds' implicit competence with their own false beliefs. *Cognition, 56*, 31–60.

Gallese, V. & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*, 493–501.

Gergely, G, Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165–193.

Goldman, A. (1989). Interpretation psychologized. *Mind and Language, 4*, 161–185.

Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences, 16*, 15–28.

Goldman, A. (2000). The mentalizing folk. In D. Sperber (Ed.), *Metarepresentations*. New York: Oxford University Press.

Goldman, A. (2001). Desire, intention, and the simulation theory. In: Malle, Moses, & Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*. MIT Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge. *Behavioral and Brain Sciences, 16*, 1–14.

Gopnik, A. & Meltzoff, A. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.

Gopnik, A. & Wellman, H. M. (1992). Why the child's theory of mind really *is* a theory. *Mind and Language, 7*, 145–171.

Grezes, J. & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions:a meta-analysis. *Hum Brain Mapp, 12* (1), 1–19.

Heal, J. (1996). Simulation and cognitive penetrability. *Mind and Language, 11*, 44–67.

Levenson, R. W. & Ruef, A. M. (1997). Physiological aspects of emotional knowledge and rapport. In W. Ickes (Ed.), *Empathic accuracy*. New York: Guilford.

Lewis, C., Freeman, N., Hagestadt, C., & Douglas, H. (1994). Narrative access and production in preschoolers' false belief reasoning. *Cognitive Development, 9*, 397–424.

Lhermitte, F., Pilon, B., & Serdaru, M. (1986). Human autonomy and the frontal lobes: I. Imitation and utilization behavior: A neuropsychological study of 75 patients. *Annals of Neurology, 19*, 326–334.

Meltzoff, A. (1999). Imitation. In Wilson and Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.

Meltzoff, A. & Moore, M. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78.

Meltzoff, A. & Moore, M. (1983). Newborn infants imitate adult facial gestures. *Child Development, 54*, 702–709.

Milner, A. D. & Goodale, M. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Mitchell, P. & Lacohee, H. (1991). Children's early understanding of false belief. *Cognition, 39*, 107–127.

Nichols, S. & Stich, S. (forthcoming). *Mindreading*. Oxford: Oxford University Press.

Nichols, S., Stich, S., Leslie, L., & Klein, D. (1996). Varieties of off-line simulation. In P. Carruthers & P. Smith (Eds.), *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Peacocke, C. (1998). Conscious attitudes, attention, and self-knowledge. In Wright, Smith, & MacDonald (Eds.), *Knowing Our Own Minds*. Oxford: Clarendon Press.

Peacocke, C. (1999). *On Being Known*. Oxford: Oxford University Press.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition, 36*, 1–16.

Pylyshyn, Z. (2000). Situating vision in the world. *Trends in Cognitive Science, 4* (5), 197–207.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141.

Rizzolatti, G., Fadiga, L., & Gallese, V. (forthcoming). From mirror neurons to imitation: Facts and speculations. In Prinz & Meltzoff (Eds.), *The Imitative Mind: Development, Evolution and Brain Bases*. New York: Cambridge University Press.

Russell, J. (1996). *Agency: Its Role in Mental Development*. Hove: Erlbaum (UK) Taylor & Francis.

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology, 9*, 331–349.

Scholl, B. J. & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4* (8), 299–309.

Stich, S. & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind and Language, 7*, 35–71.

Walton, K. (1997). Spelunking, simulation and slime: On being moved by fiction. In Hyort & Laver (Eds.), *Emotion and the Arts*. New York: Oxford University Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Woodward, A. (forthcoming). Selectivity and discrimination in infants' encoding of human behavior.

Reply to Alvin I. Goldman*

William Child
University College, Oxford

As Alvin Goldman reminds us, there are several different kinds of question at issue in debates between Theory Theory and Simulation Theory. Amongst the questions he distinguishes are these:

1. How is our competence at making first-person and third-person attributions of mental states acquired?
2. What are the concepts of mental states that children and adults deploy; what do they understand mental states to be; how do they conceptualize them?

I shall comment on claims Goldman makes about each debate.

1. Theory Theory and "non-representational" conceptions of attitudes

How do we acquire the ability to make attributions of mental states? In crude terms, Theory Theory's answer is that we acquire that ability by constructing a theory of what explains people's behaviour. So the gradual development of a child's ability to ascribe attitudes to others is explained in terms of the gradual development of its theory of others' minds. And according to what Goldman describes as "the orthodox position", part of that development is a shift "from a nonrepresentational understanding of mind to a representational understanding". Goldman has a series of objections to this position. I want to focus just on his objection to the claim that there is a stage at which children have a "purely nonrepresentational concept of desire", a concept "involving a relation between the target person's mind and some object or state-of-affairs *in the world*. Wanting is not construed by two-year-olds as involving a mental representation of

objects or states of affairs . . . [The child] can simply think of the target person as having an internal longing for an external object". Goldman's point is, in effect, that the idea of a nonrepresentational conception of wanting, or desire, is incoherent. He has two objections. But, I shall argue, neither is convincing.

Goldman's first objection is that advocates of the supposed nonrepresentational conception of desire can have no coherent story about what the relata of an unfulfilled desire could be. Against that, I shall argue that there are least two ways in which a child might make sense of someone's having an unfulfilled desire without the child's having to think of that person as having a representation of the unrealized state of affairs that she wants to be the case.

i. The child may have a primitive notion of desire that relates a person to an object, not to a state of affairs. Thus, the child may think that S wants a banana. The basis for this attribution might be that S tries to grab the banana, stares at it, resists others' attempts to get it, and so on. But the desire is unfulfilled; S hasn't got the banana. There is no obvious problem with the idea that the child can make sense of such a desire being unfulfilled without having to think of S as having a representation of how she would like the world to be. The relata of the desire are simply S and the object she desires; and the child can think of them as being related in the appropriate way whether or not S's desire is fulfilled. Admittedly, that does not give the child the resources for making sense of the idea of S desiring something that doesn't exist. But it is extremely plausible that children can make sense of desires for observed objects before they can make sense of desires for non-existent objects.

ii. A second possibility (compatible with the first) is that the child's primitive conception of desire amounts simply to the idea that S will be happy if *p*, and unhappy if *not-p*. In those terms, to think that S has an unfulfilled desire that *p* is to think that if *p* comes about then S will be pleased (or, if *p* were to come about then S would be pleased). That requires the child to be able to think about how things will be (or how things might be) and to understand the thought that things will (or might) be different from the way they currently are; in other words, the child must have representations of future or possible states of affairs. But in order to think of others' desires in this primitive way, the child need not think of S, the subject of the ascriptions, as having such representations. (Of course, a conception of desire along these lines is not a conception of desire properly so-called. Our ordinary conception of desire allows us to make perfectly good sense of the possibility that someone may be made unhappy by getting what she desires; whereas, with the primitive conception of desire that equates the idea that S desires that *p* with the idea that S will be happy if *p*, it would be a contradiction to suppose that someone might be made un-

happy by getting what she desires. That shows that we cannot understand the full-blown, adult concept of desire in terms of the idea of what will (or would) make S happy. But what is in question in Goldman's paper is not the full, adult concept of desire; the issue is simply whether there can be a *primitive conception* of desire that allows a child to make sense of the idea of an unfulfilled desire without thinking of desire as involving a representation of the desired state of affairs. The current suggestion seems to show that there can indeed be such a primitive, nonrepresentational conception of desire.)

Goldman states his second objection like this:

> It is universally agreed that understanding false belief is a sufficient indicator of a representational concept of belief. Unfulfilled desire is precisely analogous to false belief; the only difference is in the 'direction of fit' between mind and world. [So, since] 2-year-olds understand the possibility of unfulfilled desires, the parallel conclusion to draw is that 2-year-olds have a representational concept of desire.

But *is* the analogy between a false belief and an unfulfilled desire an exact one? There are two ways in which there can, in a sense, be a mismatch between desire and world: the desire may be unfulfilled; but it may also be a desire for something that is not desirable. Now it is this second notion – the idea of someone's desiring something undesirable – that is most closely analogous to the idea of false belief: in thinking that someone desires something undesirable, one thinks that she represents as good, or desirable, something that is (in one's own opinion) bad, or undesirable; and that is a close analogue to the way in which, in thinking that someone believes something false, one thinks that she represents as true something that is (in one's own opinion) false. And it is true that understanding the possibility of someone's desiring something undesirable, just like understanding the possibility of someone's believing something false, requires one to think of the subject as having representations of the world. In order to make sense of the idea that someone believes something false, I have to grasp the idea that the other person has her own perspective on the world; I must grasp the tripartite distinction between how another person believes the world to be, how I believe it to be, and how it really is.[1] The parallel for desire is that in order to be able to make sense of someone's desiring something that is not desirable, I have to grasp the tripartite distinction between what she desires, what I desire, and what is really desirable. That really is analogous to the case of false belief in involving the idea that the subject has her own perspective on the world (in this case, a perspective on what is good or desirable, or on how the world should be). But the idea of an unfulfilled desire is much simpler. One

could make perfectly good sense of the idea of someone's desiring something they have not got (and, therefore, of someone's having an unfulfilled desire) even if one were operating with a simple, undifferentiated concept of what is good or desirable – a concept that makes no distinction between what S desires, what one desires oneself and what is really desirable. Putting it that way brings out the fact that much less conceptual sophistication is required for grasping the idea of an unfulfilled desire than for grasping the idea of a desire for something that is not desirable (a desire for something that one should not desire, or that one is wrong to desire). So, it seems to me, there is no incoherence in supposing that a child might grasp the idea of unfulfilled desire before grasping the idea of false belief (or the idea of desire for things that are undesirable).

I conclude that Goldman's objections to the very idea that children might, at a certain stage, possess a nonrepresentational concept of desire are not compelling.

2. An internalist account of the attitude concepts?

In the last section of his paper, Goldman makes some speculative moves towards an account of our concept of a desire or a goal that would provide an alternative to the Theory Theorist's view that that concept is theoretical – the view that we understand the concept of desire "in terms of causal-inferential relationships to ... the environment and overt behaviour of the agent". I share his interest in identifying a sensible alternative to the Theory Theorist's view of propositional attitude concepts. But I am sceptical about the suggestion he makes.

Goldman thinks that an internalist (or partially internalist) account of attitude concepts might take either of two forms. One idea is "that occurrent attitude states have phenomenologies, and different attitude types are conceptualized in terms of their distinctive phenomenologies". Another possibility is that "there is a class of intrinsic (non-relational) features at the sub-personal level that distinguish among the attitudes. Perhaps these features are what get associated with words like 'want' and 'think'. If such features are directly detected and hence (internally) 'observed', this offers a second 'internalist' alternative to the standard Theory Theory approach". It is not clear to me whether the substance of Goldman's proposal in this paper is supposed to select between these alternatives. At any rate, if the concepts a subject deploys are components of her thoughts and beliefs, and if thoughts or beliefs are phenomena that can feature at the personal level – the level of conscious thought and experience – I

find it hard to see how an account of what it is to possess a given concept could proceed entirely at the sub-personal level.

Goldman rightly says that internalist accounts of mental concepts have been unpopular "because of [the] looming problem of other minds". But we should be clear about what problem of other minds is at issue here. We can distinguish between an epistemic/descriptive problem about other minds and a conceptual problem. The traditional epistemic problem about other minds is this: "What justifies each of us in supposing that other people have conscious states and experiences at all?" Goldman, eschewing issues of justification, substitutes a descriptive version of the question: "What *explains* our thinking that other people have conscious experiences?"

There are of course important distinctions between the epistemic version of this traditional question and Goldman's naturalized, descriptive version. But the two versions are fundamentally similar in this respect: in stating the epistemic or descriptive problem of other minds, we take it for granted that we understand *what it would be* for other people to have conscious experiences. The question is simply what justifies or explains my believing that others *do* have conscious experiences. But we can raise a question about the understanding that is simply taken for granted in posing the epistemic/descriptive problem. And that leads us to a different, and more fundamental, problem about other minds: how do I get a concept of experience that allows me to understand the thought that other people have conscious experiences; how do I so much as make sense of that thought? This is the conceptual problem of other minds. And it is this question that poses the deep problem for internalist accounts of mental concepts.

To see why, consider the concept of pain. On an internalist view, there will be two stages in explaining my acquisition of the concept of pain. At the first stage, I attach the word "pain" to a particular kind of sensation that I am introspectively aware of in my own case; at this stage, the content of the word is exhausted by the association between it and the introspected feelings of pain. At the second stage, I somehow move from the introspectively-based grasp of what it is *for me* to be in pain to an understanding of what it is *for someone else* to be in pain. But there is (to put it mildly) a real problem in seeing how this move is supposed to work. If my basic conception of pain is derived wholly from introspective awareness of my own pains, how do I ever grasp the idea of someone else's pain? In order to grasp the idea that someone else is in pain, I need to make sense of the idea of a pain that exists when I myself have no feelings of pain. But if I start with an understanding of what it is for there to be a pain that is wholly constituted by my associating the word "pain" with my

own feelings of pain, I can have no way of understanding the idea that there could be a pain even when I had no feelings of pain. On the internalist view, the thought that there is a pain is, at the first stage, the thought that there are *these*, introspectively identified, feelings; to suppose that I am not in pain is to suppose that there are no such feelings; so if my concept *pain* is explained in this purely introspective, internalist way, it is a contradiction to suppose that there might be a pain even when I feel no pain. Wittgenstein famously summed up the objection like this:

> If one has to imagine someone else's pain on the model of one's own, this is none too easy a thing to do: for I have to imagine pain which I *do not feel* on the model of the pain which I *do feel*. (*Philosophical Investigations* §302)

But the argument is not exclusive to Wittgenstein: the same basic point can, for example, be found in writings by P. F. Strawson and Bernard Williams.[2] The internalist suggests that, merely by focusing attention on our own, introspectively identified, experiences we can derive a concept of experience that applies both to ourselves and to others. The objection is that if that is the only resource we start from, we can never acquire the idea of someone else's having an experience at all. That is the fundamental problem for any internalist account of our mental concepts.

Goldman's suggestion is more subtle than I have so far suggested. For he does not propose that the whole content of any mental concept is derived by introspective awareness of one's own case. On the contrary, he insists that, in both first-person and third-person cases, there are two aspects to our mental concepts: a behavioural aspect and an internal aspect. But this subtlety in Goldman's account does not affect the point of the objection at all. For the question remains: how do we make sense of the internal aspect of the concept of a kind of mental state? In particular, how do we understand what it is for someone else's mental states to have this internal aspect? The internalist view is that our grasp of these internal aspects is achieved, in the first instance, purely by introspection in the first person case and that we then use our introspective awareness of the internal aspects of the first-person case as a basis for understanding the internal aspect in the second-person and third-person case. And at that point Goldman's proposal meets the objection that it is impossible to derive, by extension from one's own case, a concept of internal characteristics that applies to others. He cannot insulate his view from this objection simply by allowing that there is a behavioural aspect or component to our mental concepts as well as an internal aspect or component. For the key problem is a question about

the internal aspect itself; how do we make sense of the internal aspect in the third-person case?

The internalist proposal that Goldman asks us to consider is this. There is clinical and neurological evidence to support the claim that observation of the desire-manifesting, or goal-directed, behaviour of others triggers those areas of an observer's brain that are involved in executing similar movements. Suppose (to focus on the personal-level version of Goldman's suggested internalism) that this neural activity is, or gives rise to, *conscious* experience. That means that we have conscious experiences that are associated with others' goal-directed – experiences of the very same kind as those we have when we perform actions of that kind ourselves. The proposal is then that "[we] learn to match inner features detectable in [our] own goal or desire experiences with behavioural cues utilized in identifying goals or desires [in others]". If that is right, Goldman argues, then our concept of desire is not, or not wholly, a theoretical concept: it cannot be explained (wholly) in terms of a kind of behaviour or in terms of the inferred cause of a kind of behaviour.[3]

I have two objections to this theory. First, even if we ignore the conceptual problem of other minds and think only about Goldman's descriptive or explanatory question, it is very hard to see how resonance phenomena are supposed to explain why we come to believe that others' mental states have a conscious, experiential aspect. Second, once we squarely face the conceptual problem of other minds, we see that Goldman's theory has no answer to the question, how a conception derived from introspection could possibly give us a concept of experience that we can intelligibly apply to other people. But before coming to these objections, there are two preliminary points to note.

First, Goldman offers his idea about the role of mirror neurons and resonance phenomena in our grasp of mental concepts only as a speculative and sketchy suggestion. For one thing, the evidence of such resonance effects is specific to the case of desire. For a comprehensive internalist (or partially internalist) account of our concepts of mental kinds, we would need either to show that there is similar evidence of resonance for other kinds of attitude or to provide some other internalist underpinning for those concepts. One might be sceptical about the prospects for any directly analogous internalist account of the concept of belief; it is unclear how we would even begin to develop a story about observations of others' "belief-manifesting behaviour" giving rise to neural activity "characteristic of beliefs".

Second, Goldman notes that tentative suggestions in some of his previous writings that we might develop an internalist account of propositional attitudes have been taken by critics as suggesting an internalist account both of the at-

titude types and of their contents. In the current paper, he explicitly restricts the aim to that of giving an internalist account of attitude types (note 15), leaving the analysis of their contents on one side. But to the extent that his current internalist proposal is motivated by evidence about mirror neurons and resonance phenomena, it is actually not clear that he should make that restriction. The neurons that fire when goal-directed behaviour is observed are not just neurons related to action in general, but neurons related to actions of the very type observed. ("Apparently", Goldman writes, "when observing someone else perform an action, a normal human generates a 'resonant' plan, i.e., to do *the same action*, or an image *of doing it* himself" (my italics).) So, despite his explicit restriction of the intended scope of the suggestion, the logic of the position Goldman is discussing seems to push us towards an account of both attitude types and their contents.

Now, as we have seen, Goldman's version of the internalist suggestion appeals to the idea of resonance – the idea that "internal experiences of goals or desires" arise not only when we act or prepare to act ourselves, but also when we observe the goal-directed behaviour of others. But how are these resonance phenomena supposed to help?

The task is to explain how we come by a concept of desire that can be applied to others and that is not merely behavioural or theoretical but has an essentially experiential component. If our conception of others' desires had no place for introspectible experience as well as behaviour, we would be forced to choose between two unpalatable alternatives: that the concept of desire applies univocally to ourselves and others but is a purely behavioural concept; and that the concept of desire is systematically ambiguous – so that each person has a purely first-personal, experiential concept of desire that she applies to herself and a purely behavioural concept of desire that she applies to others. And neither view is tenable. We plainly have a single concept of desire that applies univocally to ourselves and others; but part of that concept is that desire has a first-person aspect – for desiring something is essentially a property of a self-conscious agent.

What we need to explain is the idea that another person's desire involves not only dispositions to behave in certain ways but also his own subjective perspective on things. But how could that feature of our concept be explained by an appeal to resonance phenomena? What resonance gives us is simply an association between the observed behaviour of others and certain of *one's own* experiences. But how does the fact that *I* have experiences of certain sorts when observing another's goal-seeking behaviour give me a grasp of the idea that *he* has experiences of some distinctive sort when he engages in such behaviour?

The facts about resonance mean that there is an introspective, experiential element to observing pain. But if that is all we can appeal to in explaining our conception of others' pain, that conception will remain purely behavioural: maybe I think of another's pain partially in terms of its experiential effects *on me*; but what is still lacking is any conception of *the other person's* experience.

The suggestion just considered is, again, less subtle than Goldman's own. For Goldman does not suggest that the mere fact of resonance is by itself enough to give us a non-behaviouristic conception of others' mental states. He writes:

> I mean to . . . proffer a possible explanation of how a child might come to interpret other people as undergoing internal experiences of desiring or planning similar to his own experiences. Such an interpretation could be facilitated by the occurrence of resonant experiences produced while observing others. After all, the observer can recognize that he does not himself act when he undergoes this experience; and in typical situations there isn't even an appropriate goal object for him to act upon. Furthermore, he sees that the person he is watching *is* acting in the appropriate way. So it might be easy for him to attribute the motoric experience he himself undergoes to the other actor. In other words, resonance activity might psychologically facilitate or promote the attribution to others of experiential (or internal) dimensions of mental states.

So the idea is not that the occurrence of resonant experiences when observing others' goal-directed behaviour is by itself *sufficient* to give us a non-behaviouristic conception of others' goal- or desire-experience; Goldman's point is just that the occurrence of such resonant experiences may *prompt us* to ascribe inner experiences to others when they behave in the relevant ways. Whether or not resonance does play such a rôle is, of course, a factual question. But, to me at least, the suggestion seems antecedently improbable. If I feel a pain every time someone else stubs his toe, why should that have any tendency at all to make me think that he feels pains when he stubs his toe? It may well make me try to stop him stubbing his toe. But, if I do not already have the idea that others have experiences just like mine, it seems unlikely that any association between his behaviour and my experiences should be enough to make me ascribe him experiences – or even to suggest that I should do so. At the very least, Goldman's suggestion needs backing up with some story about why the occurrence of resonant experiences might tend to promote the ascription of experience and sensation to others.

Furthermore, and more seriously, Goldman's suggestion still has to answer the conceptual problem of other minds. In the quotation above, resonance functions as part of an explanation of what makes us believe that others have

experiences and sensations like ours. But how are we to explain his grasp of a concept of internal experience that is applicable to others? Goldman does not explicitly address that question. But without such an explanation, the suggested internalism is vulnerable to the conceptual problem of other minds described above. Perhaps Goldman thinks that his shift from "the traditional problems of normative epistemology in the mental domain" to the non-normative problem of "finding a descriptive account of the generation of [people's] beliefs [about other minds]" neutralizes any philosophical problem of other minds. But if he *does* think this shift of perspective absolves him of the need to address the conceptual problem of other minds, he is simply wrong. The transition from justificatory questions to descriptive or explanatory questions does side-step the traditional *epistemic* problem of other minds. But, as explained above, it has no effect at all on the conceptual problem of other minds.

Goldman is right that an adequate account of our mental concepts needs to show that we apply the very same mental concepts to ourselves and to others, and that the concept of a mental state is the concept of a kind of state that both has a behavioural aspect and is distinctively available to the subject's own introspective awareness. But in developing such an account, we cannot treat the first-person case as a conceptual starting point, from which I can work outwards: first, to make sense of the idea that, in my own case, mental states have a behavioural as well as an introspective aspect; and then to make sense of the idea that others may be in states similar to my own. Focusing on my own internal experiences cannot give me the concept of a kind of state (or an aspect of a kind of state) that others can be in too. And there is no stage at which we are in a position to apply concepts of mental states (or particular aspects of mental states) to ourselves, on the basis of introspection, but not in a position to apply those very same concepts to others on the basis of their behaviour.

I conclude that the idea that Goldman floated faces a dilemma. If our conception of the "internal" aspect of mental phenomena is derived purely introspectively, then an appeal to resonant experiences will be ineffective; resonance could not put us in a position to conceive of others' mental states in a way that was not purely behavioural. But if we conceive of the internal aspect of the mental in a philosophically more satisfactory way, then we must already conceive of experience in a way that ties it to behaviour and that makes room for the idea that others as well as ourselves are subjects of experience. In that case, there is no need to appeal to resonant experiences to show how we come to grasp a conception of conscious experience that applies to others as well as ourselves.

Notes

* This paper is a revised version of a reply to Alvin Goldman's "Simulation and the grasp of mental concepts" at the CREA workshop on Mental Simulation in September 1999. I am grateful to Joëlle Proust and Jérôme Dokic, who organized the workshop, to Alvin Goldman, to the other speakers and participants for helpful comments and discussion, and to Sarah Richmond for comments on a previous draft.

1. Strictly speaking, I could make do at a rudimentary stage with just the bipartite distinction between how the other person believes the world to be and how it is – marking no distinction between my own view of the world and how things really are. But if one grasps the bipartite distinction, it is a small step to grasping the full tripartite distinction

2. See Strawson (1959: Ch. 3) and Williams (1978: 100–101).

3. Goldman's text actually offers this as a speculative suggestion about *young infants*' concepts of goal states. I hope it is not a distortion of the idea to take it also as a suggestion about how mature, adult concept-users might have a non-theoretical grasp of the concept of desire.

References

Strawson, P. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.

Williams, B. (1978). *Descartes: The project of pure enquiry*. Harmondsworth: Penguin.

Wittgenstein, L. (1958). *Philosophical Investigations* (2nd edition). G. E. M. Anscombe, R. Rhees & G. H. von Wright (Eds.), translated by G. E. M. Anscombe. Oxford: Blackwell.

Chapter 2

From simulation to theory

Paul Bernier
Université de Moncton, Canada

I. As Alvin Goldman (2000) succinctly puts it in the first sentence of his paper: "Ordinary people are mentalizers", which is to say that ordinary people are very good, naturally, at explaining and predicting the actions of others and at ascribing mental states, such as beliefs and desires, to others and to themselves. In the past several decades, philosophers and developmental psychologists have made many useful suggestions as to how we should understand mentalisation. Since the seminal paper of Wilfrid Sellars "Empiricism and the Philosophy of Mind", published in 1956, and under the influence of the work of such philosophers as David Armstrong (1968, 1980), David Lewis (1966, 1972) and others, Functionalism, or Commonsense Functionalism, has been a most influential view on the matter. The main idea of this view is that mentalisation is based on the use of a theory, that is Folk Psychology, which *posits* inner mental states as theoretical entities. Such inner states would be implicitly defined, within the theory, by their functional (or causal) role with respect to perceptual inputs, other mental states, and behavioural outputs. Thus, according to Functionalism, mentalisation requires the use of mental concepts, the mastery of which would in turn depend on an understanding, implicit though it may be, of a Functional theory, namely Folk Psychology. This idea has also had many advocates in developmental psychology, where this view has come to be known as the Theory Theory (or TT).[1]

Of course TT is not without its problems, some of which have been pressed by Alvin Goldman (1989). Indeed, in the late 1980s some philosophers and developmental psychologists have suggested different versions of what is now known as Simulation Theory (or ST), as an alternative to TT. Though different versions of ST disagree on some important points, the general idea of this approach is now very well known:[2] since I myself am an intentional agent who has particular mental states and who is disposed to act in particular ways when

I have particular mental states, and since it is very likely that I am very similar to other agents in these respects, as a matter of biological fact, then it would not be surprising if the human species had evolved a strategy by which agents use their own minds as a reliable source of information about the minds of others. According to ST, a simulator who runs a simulation of a target would use the resources of her own decision making mechanism, in an "off-line" mode, and then the mechanism would be fed with the mental states she would have if she was in the target's situation.[3] This way of stating the general idea of ST may be a bit tendentious in so far as Robert Gordon (1995a, 1995b, 1996) has formulated a version which rejects the idea that the simulator would have to figure what *she herself* would do if she was in the target's situation. On Gordon's version, the general idea can be put by saying instead that in a simulation the simulator *imaginatively transforms herself* into the target and, then, instead of actually acting on the *pretend* mental states which such a transformation yields, the simulator would explain or predict the target's actions. On this account, the simulator's focus is not directed towards herself or on how she herself would be if she was in the target's situation, but it is focused on how the world itself is supposed to be from the very perspective of the target.

Since ST has been proposed there has been many disputes between its proponents and theory-theorists, for instance as to whether ST constitutes a substantial alternative to TT, or as to which approach is most likely to be correct given some experimental data, mainly from developmental psychology and autism. But lately there has been a tendency in the literature to move towards a kind of reconciliation or, at any rate, an harmonisation between the two approaches.[4] Many have proposed so-called "mixed views" of mentalisation in which both ST and TT would explain different aspects of mentalisation. The general motivation of my paper is to help the discussion move further in the direction of such a reconciliation.

One of the early motivations of ST, as I understand it, was to move away from the over-intellectualised picture of mentalisation offered by Functionalism. Indeed, at least in the work of Gordon (1986, and especially 1995a, 1995b, and 1996) – and perhaps also in Heal (1986) – that motivation took the form of the strong claim that in order to engage in mentalisation there would be no need to master and to apply any mental concepts at all. On this way of construing simulation, an agent could be related epistemically and conatively in specific ways to her environment even if she lacks mental concepts. And it seems quite plausible that some animals, or very young children, do have beliefs and desires, even if they do not have the concepts of *belief* or *desire*. Thus understood simulation is a mechanism by which a cognitive system exploits

its own cognitive abilities in order to acquire useful information about other similar systems without requiring the mediation of mental concepts. To put it bluntly, on such a construal simulation would give us a kind of *direct* entry into the minds of others, just as it is quite plausible that as a species we evolved some perceptual systems that give us, in a certain sense, a *direct* entry into the world which we can readily exploit for our survival needs. Now, of course not all simulation theorists agree with this way of construing ST, but thus understood it would seem to be radically at odds with Functionalism. But, as I noted, my purpose in this paper is more on the conciliatory side of the issue. I want to defend a version of ST which shares the general motivation just noted, which is much in the spirit of Gordon's "radical" simulationism. But I also want to argue that this approach *must* be reconciled with Functionalism.

Now, we may wonder how a cognitive system could mentalize without mastering mental concepts. Indeed, it is quite clear that on *any* version of ST, self-ascription of mental states must play a fundamental role: the simulator must be able to reliably identify the mental sates she would have if she was in the target's place, or the mental states she has after her imaginative transformation into the target. On this basis, a theory-theorist might be tempted to argue that this requirement of reliable self-ascription constitutes a *reductio* of ST, in so far as reliable self-ascription would have to be based on the kind of behavioural and situational evidence which, on her view, is definitive of mental concepts within Folk Psychology. This problem will be at the centre of my discussion and, for brevity's sake, I will refer to it as "the self-ascription problem". Alvin Goldman (1993, 2000) has replied to this kind of concern by pointing out that it is not really a problem in so far as reliable self-ascription is not based on behavioural and situational evidence, but on phenomenal or qualitative evidence, namely on the qualitative properties of occurrent mental states which, according to him, are supposed to be accessible to introspection. Now, according to both of these views, that is TT and Goldman's introspectionist account, reliable self-ascription is seen as requiring the mastery and application of mental concepts, which is to say that both views suggest *recognitional* models of self-ascription. But, Gordon has also suggested a *non-recognitional* model of self-ascription. On Gordon's view, what he calls "ascent routines" could afford reliable self-ascriptions without requiring mastery of mental concepts. If correct, this account of self-ascription would be a quite obvious way for a simulation theorist to block the theory-theorist's objection, without having to endorse Goldman's controversial introspectionist model.

My main purpose in the rest of this paper is to articulate Gordon's solution to the self-ascription problem, namely his ascent routine model, and to defend

it against two objections that have been raised against it. I argue that the first of these objections can be resisted by modifying in a certain way Gordon's account of the ascent routine model. But, I also argue that the second objection is indeed a serious problem, and that Gordon's attempts at addressing it are inadequate. To resolve this difficulty I suggest that given the nature of ascent routines, on the understanding I develop in reply to the first objection, they are a very likely basis on which human beings can use some further cognitive capacities in order to *develop* a functional theory of mind of the kind that has been proposed by "classical" functionalists. If my argument is correct, then my view would indeed constitute a form of reconciliation between ST and Functionalism. I will conclude by briefly pointing out why Goldman's introspectionist model must be rejected.

II. According to Gordon's "radical" simulationism, while it is true that in order to simulate a target the simulator must be able to reliably self-ascribe mental states, simulation requires neither an analogical inference from the simulator to the target, nor introspection of phenomenal properties, nor the application of mental concepts. Gordon's general motivation for this austere conception of simulation is best understood on the background of the repeated attacks, in contemporary analytic philosophy, against traditional epistemological doctrines which in one way or another have given some sort of primitive or privileged status to immediate subjective experience. Now, on a certain understanding of simulation, which is shared most notably by Alvin Goldman and Paul Harris, the ascription of mental states to others (or the explanation or prediction of their actions) is mediated, first, by an ascription of mental states the simulator herself would have if she was in the target's place and, secondly, by an inference on the basis of the premise that the target is similar enough to herself, the simulator. But on Gordon's view no such analogical inference is required: within the context of simulation, the simulator imaginatively transforms herself into the target and within such a context she can *directly* self-ascribe *pretend* mental states without the mediation of an analogical inference from her to the target. Gordon's idea is that although in simulation we are limited to ascribe mental states in the *first-person*, doing so does not limit the simulator to ascribe mental states only to herself, since she can shift her "egocentric map" onto the target. In this sense the simulator would have a direct access to a perspective on the world different from her own and, thus, she would have a kind of direct access to the minds of others.

But what about the question whether self-ascription requires the application of mental concepts? In other words, how is reliable self-ascription possible

if it is based neither on behavioural and situational evidence, nor on introspection of phenomenal properties of mental states? To answer this question Gordon proposes his *non-recognitional* model of self-ascription, namely the ascent routine model, which is inspired by the following famous remark of Gareth Evans: "I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*" (1982: 225). On this view, the simulation process would not require the simulator to have second-order reflexive attitudes about her own mind, but only the ability to have first-order beliefs which are directly about external facts and not about one's mind. As Gordon notes: "Because this procedure answers a metacognitive question by answering a question at the next lower semantic level, I will call it an *ascent routine*" (1995b: 60).

According to Gordon, ascent routines can afford reliable self-ascription by linking the ascription of a mental predicate with the verbal *expression* of the mental state thus expressed. For example, a subject normally expresses her belief that p by asserting "p", and the ascent routine would allow her to ascent to "I believe that p" simply by prefixing her normal expression of the belief with the predicate "I believe". Gordon suggests that this is how children learn to use mental predicates. The important point to note is that in order to do that the child does not need to understand the concept of the mental state, all she needs is to be capable of having the mental state and of verbally expressing it. Gordon also finds support for this view in the fact that it happens often that we use linguistic forms like "I want x", or "I believe that p", not to report (or to describe) one of our own mental states, but simply to express the mental state.

It should be noted that this notion of self-ascription is quite minimal. And indeed Gordon makes it clear that ascent routines, on their own, are insufficient to yield genuine comprehending self-ascriptions of mental states. And, as it has been pointed out by Goldman (2000), it remains relatively unclear how, on this view, we should understand genuine comprehending self-ascriptions. Though we must certainly agree with Gordon that in many contexts we can use sentences of the form "I want x" or "I believe that p", to express and not to report (or describe) our own mental states, it should be granted that such linguistic forms can also be used to report (or to describe) our own mental states, which reports presuppose the understanding of mental concepts. But, it is unclear how Gordon's view can accommodate such genuine comprehending self-ascriptions. This is one of the two problems I want to address in what follows. But to do so it will be helpful, first, to focus on another difficulty which has been raised by Goldman (2000).

III. As Goldman (2000) notes it: "how can [the ascent routine] account for self-attributions of attitudes other than belief?" Take as an example of such an attitude John's desire to eat, or John's desire that he himself have something to eat. Now, if we take Evans's famous quote as a kind of recipe for the characterisation of the ascent routine, then radical simulationism would indeed seem to be in trouble. To wit, how could we fill the following blank: "I get myself in a position to answer the question whether I desire that *p* by putting into operation whatever procedure I have for answering the question whether ____". And of course answering whether p is the case would be of little relevance to answer the question whether I *desire* that *p*. But, as I want to argue, this concern arises only because of the temptation to understand the general idea of the ascent routine exclusively on the basis of this famous quotation from Evans.

On my understanding, the central idea of the ascent routine is the following. It is quite plausible that mental states, or at any rate mental states such as beliefs and desires which are at the core of mentalisation, have some natural ways of being expressed in various contexts. It is also plausible that as young children we learn to associate mental predicates that have currency in our mother tongue with their various natural expressions. Such natural expressions involve first typical gestures, and typical linguistic (or proto-linguistic) behaviours. For example, a young child who has learned some basic uses of the word "cat" may utter "cat" to naturally express her belief that there is a cat in the vicinity. But sometimes she may utter "cat", perhaps in a different tone of voice, to *inquire* whether the cat is in the vicinity, or yet in another tone of voice to *express her desire* that the cat be in the vicinity. And it seems indeed very plausible that in order to be able to do that the child needs no *mental* concepts.

Now to get back to the question raised by Goldman, it seems to me that to answer the question "Do you desire that the cat be in your arms?", the child does not need to recognise that she has a certain mental state falling under a certain mental concept, whether it is characterised by its functional role or by some typical phenomenal property. On the ascent routine model, asking the question to the child is simply asking her to *express*, or to express once again, her mental attitude. If indeed it is a desire, not a belief, nor an inquiring attitude that she has, then it is likely that she will answer the question by a further, perhaps more insisting, *expression* of her attitude. Thus it would seem that in order to address the problem of self-ascription of mental states other than belief, the radical simulationist would do well to generalise Evans's remark in such a way that it can apply to all attitudes. And to do so I suggest the following, (where "PA" is schematic for any verb of propositional attitude): "I get myself in a position to answer the question whether I PA that *p* by putting into operation

whatever procedure I have at my disposal, in context, to *express* my current mental attitude towards *p*". The important point to note about this suggestion is that in order to be capable of expressing one's current mental attitude towards *p* one needs not master the concept of that attitude. All that one needs, and this is important, is to master whatever concepts are required in order to entertain the proposition *that p*. In other words, on this understanding of the ascent routine it is true that some intentional, or representational, capacity is required, and thus some conceptual mastery, but this capacity need not involve the mastery of *mental* concepts. We should also note that in this formulation the suggestion is still restricted only to *propositional* attitudes, or intentional states. It is not totally clear to me whether this move could be generalised to apply to all mental states including sensory states such as colour experiences and raw feels. But for my purpose in this paper, I will be content with the more restricted suggestion to generalise the idea to all *propositional* attitudes. Thus, we should at least agree that this alleged problem of self-ascription of propositional attitudes other than belief is hardly a problem for the ascent routine model.

One might be tempted to object that in order to put into operation whatever procedure is required to express one's current mental attitude towards *p*, the subject would need some sort of internal phenomenal cue, in order to reliably identify the present mental state she is in (perhaps to figure out what she should do next). But, in reply to this objection, we only need to insist that on the understanding of the ascent routine I have suggested, the crucial link that is established is not between a mental predicate and some introspected qualitative property, but between a mental predicate and the natural *expression* of a mental state, which does not require the mastery of mental concepts. It is rather unclear why the expression of one's mental states should be mediated by the introspection of some internal phenomenal cue, as if one would first need to correctly identify the mental state in order to be capable of expressing it. On the ascent routine model, as I understand it, it is not a putative capacity to reliably identify a mental state, independently of our natural capacity to express it, which allows its reliable expression, but it is the naturally reliable unmediated *expression* of the mental state which is constitutive of its reliable identification.

One might also be tempted to object that not all mental states are expressed, and that it is specifically in the case of unexpressed mental states that their qualitative features is crucial to self-ascription. I think that this point is well taken, but it is a much weaker claim than the general claim that the identification of some phenomenal cue is needed in order to express the state. Moreover, it seems that unexpressed mental states are whatever states they are rela-

tive to how we otherwise would express them. I am even tempted to speculate that if unexpressed desires and beliefs have any phenomenal properties at all, it is quite plausible that these are properties of the motor images of their natural expressions, which would indicate that reliable identification of unexpressed mental states also depends essentially on links with their natural expression.[5]

For these reasons, I do not think that the radical simulationist should be too concerned with the problem of self-ascription of attitudes other than belief. It would seem that this is not really a problem, but only a consequence of the too specific formulation of the view given by the famous quote from Evans. So let me now turn to the second problem I noted above: the problem of genuine comprehending self-ascription.

IV. As I already noted, even if we can agree with Gordon that in many contexts we can use sentences of the form "I want x" or "I believe that p", to express and not to report (or describe) our own mental states, it seems obvious that such linguistic forms are also used in some contexts to report (or describe) our own mental states, in which cases we do apply mental concepts. Indeed Gordon is quite explicit in denying that in and of itself the ascent routine model would provide an account of such genuine comprehending self-ascriptions. But if ascriptions of mental states, whether in the first- or in the third-person case, is to be understood within the general framework of ST, and if moreover it is to be understood as denying that such ascriptions are based on an implicit theory of mind, then it remains unclear how genuine self-ascriptions should be understood on such a view. Of course, Gordon is not denying that as mature adults we do master and use mental concepts, and that we are capable of having second-order reflexive attitudes about our mental lives. But the difficulty is to understand how radical simulationism relates to such reflexive capacities which do involve application of mental concepts. Gordon seems to be aware of this difficulty, and he does make some attempts to address it. But this is where I part company with him, since I find his suggestions inadequate. As Goldman (2000) points it out, Gordon's suggestion is that "ascent routines yield such comprehension only when they are embedded within a simulation". According to Gordon the use of ascent routines *within the context of simulation* would suffice to yield, eventually, genuine comprehending ascriptions of mental states, and Goldman finds it hard to understand how the two elements, that is the ascent routine and simulation, combine to yield genuine, comprehending self-ascription. Here I must say that I share Goldman's scepticism. In a nutshell, my suggestion is that a radical simulationist should simply concede that in order to get genuine comprehending ascription the use of ascent routines *within*

the context of simulation must be supplemented by some other cognitive mechanism, most probably the kind of mechanism by which we develop theories about the world, be they only folk theories like folk physics for instance. In other words, my suggestion is that to achieve genuine comprehending ascription, we must develop a theory of mind, but this partly on the basis of the use of ascent routines *within the context of simulation*. In fact, if ascent routines should be characterised along the lines I have suggested, namely as simple procedures by which we associate mental predicates with the natural expression of mental states, then it would seem quite natural that the application of this kind of procedure, together with some further cognitive mechanism which enables us to rationally organise our experience, would yield precisely the kind of functional theory of mind which has been suggested in "classical" Functionalism. But before developing and defending this claim let me first say a bit more about why Gordon thinks that ascent routines *within the context of simulation* could, on their own, yield genuine comprehending ascriptions, and why I think this suggestion is inadequate.

Here is how Gordon states the point:

> [...] suppose that in simulating another [...] – in describing "the world" as it is within the context of the simulation – [the simulator] makes some assertions that contradict assertions of her own, that is, those she makes from her "home" perspective. In other words, she allows "the facts" within a simulation to deviate from the actual facts. And suppose further that she introduces these deviations, not in an arbitrary way, but in response to situational and behavioural evidence, in a way that reduces error in explanation and prediction. (1997b: 61)

Gordon then goes on to suggest that by applying the ascent routine the simulator would be making "motivated attributions of false belief" (*Ibid.*). The least that can be said is that this way of addressing the issue is a bit sketchy, and it is hard to see how the understanding of the concept of belief is supposed to follow simply from self-ascription, within the context of simulation, of beliefs which conflict with one's "home" beliefs. Of course, the suggestion hints in the direction of the necessity to acquire a sense that the world is being viewed from certain perspectives. And here I think that Gordon is right about this. But he goes further, and he suggests that this capacity to self-ascribe, *in the context of simulation*, beliefs that are in conflict with one's "home" beliefs would suffice for what he calls "a routine proficiency in using the concept of belief" (*Ibid.*). But here I simply do not see how this is supposed to follow.

The main reason why I do not find Gordon's proposed solution adequate is that it is totally unclear how it is supposed to account for genuine comprehending self-ascription *in one's own case*, that is self-ascription with the understanding that the belief one is self-ascribing could be false. Suppose I use the ascent routine to ascribe to myself the belief *that p*, and suppose I also use the ascent routine, but this time in the context of a simulation, to ascribe to a target T the belief *that p′*, where *p* and *p′* are logically inconsistent. How is this supposed to be sufficient to make my self-ascription of the belief *that p* a comprehending one? How is this supposed to help me understand that one of the two beliefs must be false and that it could very well be mine? Moreover, how am I to recognise that there is an inconsistency between the two beliefs? It is difficult to see how I could understand these things unless I understood much more, for instance that both beliefs aim at describing one and the same objective reality and that two beliefs that are logically inconsistent can hardly do both these things at once.

Now, I am inclined to think that what is missing from radical simulationism is precisely what the subject would get by acquiring a functional theory of mind, that is a theory which *posits* inner mental states as theoretical entities, which are implicitly defined, within the theory, by their functional (or causal) role with respect to situational and behavioural evidence. Whatever the functional theory, we should expect that it rules out that the belief *that p* and the belief *that p′* could have the same functional role. In fact Gordon's own attempted answer to the problem, in the passage I just quoted, even hints in that direction, when he notes that the simulator "introduces deviations from [her own beliefs], not in an arbitrary way, but in response to situational and behavioural evidence, in a way that reduces error in explanation and prediction" (*Ibid.*). Introducing such deviations in a non arbitrary way is just what we should expect the simulator to do in the process of developing a functional theory of mind.

It should be noted, however, that in the passage I quoted above, Gordon's purpose is only to support the claim that ascent routine *within the context of simulation* are sufficient to yield "a routine proficiency in using the concept of belief" (*ibid.*). But, he also notes that this is not sufficient to yield a *sophisticated* understanding of the concept of belief, and he makes an intricate suggestion about how a subject could come to acquire such a *sophisticated* understanding, that is by *simulating herself from an alien perspective:*

> To see her own present beliefs as distinguishable from the facts she will have to simulate another for whom the facts are different – or more broadly adopt

> a perspective from which the facts are different, whether this perspective is occupied by a real person or not – and then from the alien perspective, *simulate herself*. Or at least she will have to realise abstractly that from a perspective different from her own present perspective what she now counts as fact may be, not fact, but nothing more than fact-from-her-particular-perspective. (1995b: 62)

It is unclear to me what we should make of this suggestion. Be that as it may, it certainly invites a reading on which the subject would objectify herself, so to speak, and would use whatever behavioural and situational evidence is relevant to simulate herself, just as she would in the case of simulating someone else. But it seems quite implausible that when we ascribe beliefs to ourselves, *with an understanding of the concept of belief*, we normally do so on the basis of behavioural and situational evidence. It seems much more plausible that we self-ascribe the belief directly, by applying the ascent routine, and that insofar as we have acquired a functional theory of mind, it is this theory which gives us the understanding of the concept we self-ascribe, allowing us to say a sentence of the form "I PA that p", not to express one's PA-ing that p, but to report (or describe) it, in a conceptually competent way.[6]

Gordon's central idea seems to be that the regular use of ascent routines would enable the subject to shift perspectives. And this kind of shift in perspectives would introduce a certain complexity which would allow the subject to *understand*, for example, why an utterance of "p, but I do not believe that p" though infelicitous is not strictly speaking inconsistent, since one could make sense of the two parts of this utterance, namely "p" and "I do not believe that p", as being made from within different perspectives. And this would be the key to a proper understanding of the concept of belief. I think we can hardly deny that understanding that beliefs – and other mental states – are held from a certain perspective is an essential element in mastering the concept of belief, but it remains unclear why this would suffice to yield genuine comprehending self-ascription. But suppose that self-ascription, whether in the "home" case or within the context of simulation, automatically proceeds on the basis of an ascent routine, namely a routine by which we automatically associate a mental predicate with the natural expression of a mental state. Then it would be quite plausible that a subject who regularly uses such ascent routines, given that she also has other cognitive capacities to understand regularities in the world, could come to realise that similar natural expressions occur under similar circumstances, in one's own case as well as in the case of others. And, it also seems quite plausible that a subject who has other general cognitive capacities to organise the objective world, would hypothesise that there is a common in-

ner cause of such expressions. That is: it would be quite natural for the subject – you may want to call her Jones, or one of our Rylean ancestors[7] – to hypothesise that some typical human expressions are caused by some typical inner *posits*, which are in turn produced in similar typical evidential circumstances. Does this imply that in genuine comprehending self-ascription the subject would have to simulate herself from an alien perspective? Or that she would have to ascribe the mental states on the basis of behavioural and situational evidence? I don't think so. The subject could still self-ascribe the mental state directly by the use of an ascent routine, but this time she would be in a position to understand that the mental predicate she uses denotes a mental concept which is defined within her newly acquired functional theory, a theory which subsumes any individual to which ascent routines can be applied, that is, herself as well as any *pretend* subject which are referred to in the context of a simulation.

For these reasons, I think that if we take seriously the ascent routine model of self-ascription, then the best way to solve the problem of genuine comprehending self-ascription is to hold that the use of ascent routines is precisely what enables us, given some further cognitive capacities, to develop a functional theory of mind. Such a wedding between radical simulationism and "classical" Functionalism is a bit ironic. In fact if we accept it, then radical simulationism would be anything but a radical departure from the orthodox functionalist view. Indeed, if this understanding of radical simulationism is correct, then it is perhaps Goldman's version of ST which should be labelled as its "radical" version, in so far as it proposes an account of self-ascription which involves introspection of some phenomenal properties, a view which is foreign not only to "classical" Functionalism but also to Gordon's version of ST. Let me conclude very briefly by saying why I find Goldman's introspectionist approach difficult to accept.

V. As I noted earlier, the ascent routine model is not the only way to address the problem of self-ascription, from the point of view of ST. Alvin Goldman (1993, 2000) has suggested an introspectionist model of self-ascription according to which a subject would reliably identify her own mental states by using "information about the *intrinsic* (nonrelational) and *categorical* (nondispositional) properties of the target [mental state]" (1993: 87). Moreover, on this view, these intrinsic nondispositional properties are qualitative properties of the target mental states. But, as it has been pointed out by Shaun Nichols and Stephen Stich, Goldman's proposal can be interpreted in two different ways.[8] In fact, Goldman's suggestion seems to be restricted to the more limited, or weak, claim that such a recognitional mechanism would suffice to discriminate

and classify various *kinds* of mental states. In other words, the system could tell beliefs apart from desires, hopes, and other kinds of mental states on the basis of what it feels like to occurrently have a belief, as opposed to, what it feels like to have say desires, hopes, and so on. But this weak introspectionist claim is quite obviously insufficient to account for how a subject comes to know what particular intentional state she has. After all, it seems hard to deny that the content of a particular intentional state is essential to its identity. In other words, the view is insufficient to explain how a subject can reliably self-ascribe mental states, in a comprehending way, or how she knows that she has the belief *that p* and not a belief *that p′*. But certainly, an adequate account of reliable self-ascription should answer this question. Thus this weak interpretation of Goldman's introspectionist view is clearly inadequate. Here the appeal of the ascent routine model is quite obvious: indeed we should expect that a desire for a glass of milk is naturally expressed in a way which is significantly different from the way a desire for a walk in the park is normally expressed, while it is an open question what qualitative properties might distinguish these two mental states.

Goldman might be tempted to make his introspectionist model stronger by defending the claim that it can also account for how a subject reliably self-ascribe particular intentional states. But then the view becomes extremely difficult to believe, because it's hard to see why we should accept it unless it would be developed into a full blown theory of content based essentially on the putative qualitative aspects of content, and this is highly controversial. Moreover, given the plausibility of semantic compositionality, productivity, and systematicity, it would seem that the system responsible for the classification of mental states on the basis of their content's putative qualitative properties would have to be capable of making a tremendous amount of qualitative discriminations, which would seem to be an excessive demand on a cognitive system. At any rate, it should be clear that such a strong interpretation of Goldman's introspectionist view has some very far reaching consequences which seem *prima facie* difficult to accept, mainly that a theory of meaning would have to be formulated in terms of a theory of qualitative properties.

To sum up, in this paper I have argued that simulation theory, on any of its versions, must address what I have called the self-ascription problem, given the crucial role of self-ascription in simulation. I have defended Gordon's ascent routine model against what I have called the problem of ascribing attitudes other than belief, arguing that it is not really a problem, since Evans's suggestion can quite easily be generalised, once we make it clear that the ascent routine links mental predicates to different types of *expressions* of various

mental states, and that doing so does not entail that the subject would have to master mental concepts, but only that she would have to master the concepts involved in the mental content. I have also argued that the problem of genuine comprehending self-ascription is indeed a problem for radical simulationism, and a problem that Gordon has failed to address adequately. I have argued that this problem can readily be solved once we realise that given the nature of ascent routines, it is extremely plausible that our regular application of such routines, together with some other cognitive capacities, would naturally lead us to acquire a functional theory mind of the kind that has been proposed in "classical" Functionalism. On the picture I have suggested, simulation is what gives us a *direct* entry into the minds of others, by the use of ascent routines *within the context of simulation*. This is, so to speak, how we observe the minds of others, namely how we observe the world from a perspective different from one's own, by applying the ascent routine within the context of simulation. But having a direct entry into the minds of others is hardly sufficient for reflexive understanding. Just as sensory stimulation gives us a direct entry into the world, it is hardly sufficient to give us a fine-grained and rationally coherent understanding of it. To that end some theory-building is required.[9]

Notes

1. See Leslie (1987, 1994), Leslie and German (1995), Perner (1991, 1996), Gopnik and Wellman (1995), Metlzoff and Moore (1995), Gopnik and Meltzoff (1997).

2. For different versions of ST, see Heal (1986, 1996a, 1996b), Gordon (1986, 1995a, 1995b, 1996), Goldman (1989, 1995, 2000), Harris (1995).

3. See Stich and Nichols (1992) for a detailed formulation of ST along these lines.

4. See for instance Heal (1996a), Perner (1996), Perner et al. (1999), and Stich and Nichols (1997).

5. On motor imagery, see Jeannerod (1994).

6. In fairness to Gordon, I should note that in personal communication he has indicated that this reading of what he intended to say by talking of simulation of oneself from an alien perspective was a misunderstanding of his view. But he agreed that the passage I just quoted does invite such an interpretation, which might explain why he seems to be shying away from this idea of simulating oneself from an alien perspective. Be that as it may, it still remains unclear how the use of ascent routines within the context of simulation would yield genuine comprehending self-ascription.

7. See Sellars (1956).

8. This objection has been raised by Nichols and Stich in "Reading One's Own Mind: A Cognitive Theory of Self-Awareness".

9. I thank the participants to the Conference *Simulation and Understanding of Action*, where a version of this paper was presented. I owe special thanks to Joëlle Proust and Jérôme Dokic, and to Gianfranco Soldati for his reply. I also thank John Campbell, Alvin Goldman, Robert Gordon, Daniel Laurier, Donald Peterson, and Stephen Stich.

References

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

Armstrong, D. M. (1980). The Causal Theory of the Mind. In *The Nature of Mind and Other Essays*. University of Queensland Press.

Davies, M. & Stone, T. (Eds.). (1995a). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.

Davies, M. & Stone, T. (Eds.). (1995b). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Goldman, A. I. (1989). Interpretation Psychologized. *Mind and Language, 4*, 161–185. Reprinted in M. Davies & T. Stone (Eds.), 1995a (pp. 74–99).

Goldman, A. I. (1993). The Psychology of Folk Psychology. *Behavioral and Brain Sciences, 16*, 15–28.

Goldman, A. I. (1995). In Defense of the Simulation Theory. In M. Davies & T. Stone (Eds.), 1995a (pp. 191–206).

Goldman, A. I. (2000). The Mentalizing Folk. In D. Sperber (Ed.), *Metarepresentation*. Oxford: Oxford University Press.

Gopnik, A. & Meltzoff, N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gopnik, A. & Wellman, H. M. (1995). Why the Child's Theory of Mind Really *Is* a Theory. In M. Davies & T. Stone (Eds.), 1995a (pp. 232–258).

Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind and Language, 1*, 158–171. Reprinted in M. Davies & T. Stone (Eds.), 1995a (pp. 60–73).

Gordon, R. M. (1995a). The Simulation Theory: Objections and Misconceptions. In M. Davies & T. Stone (Eds.), 1995a (pp. 100–122).

Gordon, R. M. (1995b). Simulation Without Introspection or Inference from Me to You. In M. Davies & T. Stone (Eds.), 1995b (pp. 53–67).

Gordon, R. M. (1996). 'Radical' Simulationism. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 11–21). Cambridge: Cambridge University Press.

Harris, P. L. (1995). From Simulation to Folk Psychology: The Case for Development. In M. Davies & T. Stone (Eds.), 1995a (pp. 207–231).

Heal, J. (1986). Replication and Functionalism. In J. Butterfield (Ed.), *Language, Mind, and Logic*. Cambridge: Cambridge University Press. Reprinted in M. Davies & T. Stone (Eds.), 1995a (pp. 45–59).

Heal, J. (1996a). Simulation and Cognitive Penetrability. *Mind and Language, 11*, 44–67.

Heal, J. (1996b). Simulation, Theory and Content. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 75–89). Cambridge: Cambridge University Press.

Jeannerod, M. (1994). The Representing Brain: Neural Correlates of Motor Intention and Imagery. *Behavioral and Brain Sciences, 17*, 187–246.

Leslie, A. M. (1987). Pretense and Representation: The Origins of 'Theory of Mind'. *Psychological Review, 94*, 412–426.

Leslie, A. M. (1994). Pretending and Believing: Issues in the Theory of ToMM. *Cognition, 50*, 211–238.

Leslie, A. M. & German, T. P. (1995). Knowledge and Ability in Theory of Mind. In M. Davies & T. Stone (Eds.), 1995b (pp. 123–150).

Lewis, D. (1966). An Argument for the Identity Theory. *Journal of Philosophy, 63*, 17–25.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy, 50*, 248–259.

Metlzoff, A. N & Moore, M. K. (1995). Infants' Understanding of People and Things: From Body Imitation to Folk Psychology. In J. L. Bermúdez, A. Marcel, & and N. Eilan (Eds.), *The Body and the Self* (pp. 43–69). Cambridge, MA: MIT Press.

Nichols, S. & Stich, S. (forthcoming). Reading One's Own Mind: A Cognitive Theory of Self-Awareness. In Q. Smith & A. Jokic (Eds.), *Aspects of Consciousness*. Oxford: Oxford University Press.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Perner, J. (1996). Simulation as Explicitation of Predication-Implicit Knowledge about the Mind: Arguments for a Simulation-Theory Mix. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 90–104). Cambridge: Cambridge University Press.

Perner, J., Gschaider, A., Kühberger, A., & Schrofner, S. (1999). Predicting Others Through Simulation or by Theory? A Method to Decide. *Mind and Language, 14*, 57–79.

Quine, W. V. O. (1996). Progress on Two Fronts. *Journal of Philosophy, 93*, 159–163.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In H. Feigl & M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science*, vol. 1 (pp. 309–330). Minnesota: University of Minnesota Press.

Shoemaker, S. (1993). Special Access Lies Down with Theory Theory. *Behavioral and Brain Sciences, 16*, 78–79.

Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Stich, S. P. & Nichols, S. (1992). Folk Psychology: Simulation or Tacit Theory? *Mind and Language*, 7, 29–65. Reprinted in M. Davies & T. Stone (Eds.), 1995a (pp. 123–158).

Stich, S. P. & Nichols, S. (1997). Cognitive Penetrability, Rationality and Restricted Simulation. *Mind and Language, 12*, 297–326.

Reply to Paul Bernier

Gianfranco Soldati
University of Freiburg, Switzerland

I. One reason why some people have found simulation theory (ST) attractive is that it appears to make it possible for a subject to understand somebody else's mental life without having to use a theory about how mental life in general is supposed to work.

The problem of course is what it means *to understand* somebody else's mental life. If understanding presupposes the capacity to use concepts, and if possession of a concept requires possession of a theory, then ST, as described above, cannot even get started. Thus, in order to consider ST as a serious proposal, we should either find a notion of *understanding* which does not involve possession of concepts or a notion of concept possession which does not presuppose knowledge of a theory.

One central argument in Bernier's paper is that in order to yield a genuine understanding of somebody else's mental states, one ought to have a full grasp of mental concepts; and a full grasp of those concepts involves knowledge of a particular theory, namely functionalism. Bernier, however, thinks that this view can, or rather should be, accommodated with a form of radical simulationism. Given my initial remarks, this should certainly appear surprising, of which Bernier himself is well aware.

II. Let me then first summarize Bernier's argument before I try to explain where I think the problems are.

Bernier accepts Gordon's radical simulationism as far as the two following points are concerned:

– In the context of simulation, the interpreting subject "imaginatively transforms herself into the target and ... *directly* self-ascribes *pretend* mental states without the mediation of an analogical inference from her to the target";

– The *minimal* self-ascription involved in simulation does not entail a grasp of mental concepts, since it uses the *ascent routine*: "I get myself in a position to answer the question whether I PA [where "PA" is schematic for any verb of propositional attitude] that p by putting into operation whatever procedure I have at my disposal, in context, to *express* my current mental attitude towards p". So, when I ask my three-year-old son 'what do you believe you are holding in your hand?' he says: 'ice cream' in a way I would hardly confuse with his demand 'ice cream' when asked to say what he wants me to buy. Thus he should be able to use those expressive procedures if asked to simulate my mental states when I believe to be holding an ice cream and when I wish somebody to buy one for me.

It is agreed that the simulation mechanism described so far does not require the child to grasp the mental concept under consideration. The child does not need to know what a desire is in order to express it. And he does not need to make a judgement involving the concept DESIRE in order to simulate it.

Bernier thinks that this approach faces a serious problem, the problem of the *genuine comprehending self-ascription*. Now, what is a *comprehending self-ascription*? The most obvious answer is that such a self-ascription is a full-fledged conceptual ascription. It should be clear by now that the minimal self-ascription involved in radical simulation does not satisfy, indeed it is not meant to satisfy, that requirement. So this fact alone can hardly generate a problem.

One would like to be told, however, how the subject is supposed to move from the minimal self-ascription to the genuine comprehending self-ascription. This appears to be a question concerning concept acquisition; and one would certainly want to know whether the simulation process described above plays any role in it, if any at all. The point at this stage could be that the simulation process has to be completed with some form of theorizing in order to yield comprehending self-ascription. But again, this point need not worry the radical simulation theorist, at least not as long as he does not commit himself to a specific account of concept acquisition. For one, one might wonder whether concept acquisition generally requires a theoretical framework. I shall return to this point in a moment.

It may be felt that minimal self-ascription, as described above, does not suffice to yield a *real understanding* of the target's mental states. It might be required, for instance, that in order to simulate the target's mental states, the subject ought to have a particular sensitivity to contradictions occurring within his own system of self-ascriptions. This makes sense since the subject ought to be somehow sensitive to the difference between the actual *home facts* and the

facts within the context of simulation. Given a belief that p, and a contradicting belief that p^*, the subject ought to be able to make a difference between self-attributing the belief that p^* in the actual mode and attributing it in the simulation context as a belief he would have if the target's context would obtain. Bernier states that in order to make such a difference the subject ought to have understood that the two contradicting beliefs "aim at describing one and the same reality and that two beliefs that are logically inconsistent can hardly both do that at once" (13). I am not sure I see why this should be so. Can a child not experience a contradiction without having the concept BELIEF? What feature of the simulation context is supposed to generate the need for the child to have the nearly philosophical concept of a belief describing reality? Some more details would certainly be necessary here.

III. Bernier makes a positive suggestion as far as the integration of radical simulation into functionalism is concerned. If I understood it correctly, his story proceeds as follows:

1. Via the ascent routine the subject *directly* associates a natural expression to an experienced mental state. This association is supposed to be direct in the sense that it is not recognitional, it does not rely on any evidence, either behavioral or phenomenal.
2. In a second phase the subject comes to "realize that similar natural expressions occur under similar circumstances, in one's own case as well as in the case of others" (15).
3. The subject then comes to "hypothesize that there is a common inner cause to such expressions". The subject thus develops the idea that "some typical human actions are caused by some typical inner posits".
4. The subject attains a full grasp of the mental concepts applied to him- or herself through an understanding of the fact that "the concept she applies to herself is the same as the one she applies to others, the one which is defined by whatever functional role it has in his functional theory of the mind".

As Bernier puts it at the end of his paper, the ascent routine gives us a direct entry into the minds of others, but having a direct entry is hardly sufficient for understanding, just as our direct sensory entry into the world is hardly sufficient to understand it.

Now, I take it that the account under consideration is meant to show how the capacity to make judgements involving mental concepts hooks up with the ability to make simulated minimal self-ascriptions. If this is the goal, it

is not enough to show that the capacity involved in the ascent routine needs to be supplemented with a functional theory in order to yield comprehensive self-attributions. It should also become clear that full-fledged conceptual self-ascriptions of mental predicates somehow depend on the ability to get involved in simulated minimal self-ascriptions. This is precisely what I am unable to see. Indeed, traditional functionalist theories typically do not involve any reference to minimal self-ascription. The capacity of a simulated minimal self-ascription, furthermore, does not seem of any relevance for the determination of the mental concepts involved in comprehensive self-ascriptions. Take for instance steps 2 and 3, as described above: "the subject realizes that similar natural expressions occur under similar circumstances, in one's own case as well as in the case of others ... and eventually comes to make the hypothesis that there is a common inner cause to such expressions". Here it sounds as if the subject were observing himself as one among many others and formulating generalizations on the basis of his observations. Why should the subject be engaged in any simulating self-ascription in order to do so?

In the light of these considerations it would appear that for all we know, somebody who lacks the capacity to get involved in minimal simulated self-ascription could, at least in principle, possess the very mental concepts we use in our attributive practice. This, however, gives us a very weak form of ST: our distinctive capacity of obtaining a direct entry into other minds would be irrelevant for our understanding of somebody else's mental life. I wonder whether this very weak ST is not a consequence of the fact that we are asked to make it compatible with functionalism.

Chapter 3

Neurophysiological evidence for simulation of action

Jean Decety
INSERM, Lyon and University of Washington, Seattle

What do we mean by action?

In recent years, the domain of the physiology of action has been studied at many different levels, from single cell recording in monkey to functional imaging in healthy human volunteers. More importantly, the revival of interest in action may be seen as the consequence of the rapid growth of cognitive neuroscience, which is an interdisciplinary melding of studies of the brain, of behavior and cognition, and of computational systems. Not only each approach constrains the others, but rather each approach provides insights into different aspects of the same phenomena. In this heuristic perspective, information processing theory is not separated or independent of the properties of the neural substrate.

An action may be described as the outcome of several information processing stages, i.e. intention, planning, preparing and execution. According to the Causal Theory of Actions (e.g. Davidson 1991; Searle 1983) what distinguishes actions from mere happenings is the nature of their causal antecedents. Indeed, a goal-directed action is often internally generated. This implies that the generation of action involves a representational stage which is, here, synonymous of mental (or cognitive) representation. However, it is clear that the concept of mental representation of action designates both the mental content related to the goal or the consequences of a given action and the covert neural operations that supposed to occur before an action begins. There is no ontological reason to consider these two levels of description as separate and, least of all, independent from one another. Rather, contemporary approaches to this domain

attempt to bridge with success both the cognitive description of representation of action and the neural circuitry subserving it.

The idea that perception and action are intimately linked is not new. Indeed, Sperry proposed in 1952, that perception is basically an implicit preparation to respond, and even more, its function is to prepare the organism for adaptive action. Gibson (1979) stressed that not all organisms pick up the same information from the environment, but rather they would resonate with information that is coordinated with their own potential for action. In fact, the study of perception and action is fundamentally the study of sensorimotor transformations. Sensory and motor data are generally in different formats and may refer to the same entities but in different coordinate systems.

There are several contemporary models that attempt to characterize the linkage between perception and action. For instance, Prinz (1997) has put forward the idea of a common coding model which postulates that perceived events and planned actions share a common representational domain. This model assumes: (1) that event codes and action codes should be considered as the functional basis of percepts and action plans, respectively, and; (2) that they share the same representational domain and are therefore commensurate. There is indeed supporting evidence from induction paradigms (i.e. how certain stimuli induce certain actions by virtue of similarity) and interference paradigms (i.e. mutual interference between the perception of ongoing events and the preparation and control of ongoing action) that are compatible with this model (see Prinz 1992). Some researchers in the domain of imitation have offered similar theoretical models to account for the mechanism underlying early imitative behavior. Indeed, Meltzoff and Moore (1999) suggest that infant imitation depends on a process of active intermodal mapping between what she sees and what she feels.

Another approach has been developed in the field of computational motor control. A useful conceptualization was given by Jordan (1994) for whom there are two basic kinds of transformations that can be considered to manage the relationships between sensory variables and motor variables: sensory-to-motor transformations and motor-to-sensory transformations. The transformation from motor variables to sensory variables is accomplished by the environment and the musculoskeletal system; these physical systems transform efferent motor actions into reafferent sensory feedback. It is also possible to consider internal transformation, implemented by neural circuitry, that mimics the external motor-to-sensory transformation. Such internal transformations, or internal forward models, allow the motor control system to predict the consequences of particular choices of motor actions. We can also consider internal models

that perform a transformation in the opposite direction, from sensory variables to motor variables. Such transformations are known as internal inverse models, and they allow the motor control system to transform desired sensory consequences into motor actions that yield these consequences. Internal inverse models are the basic module in open-loop control systems. Thus, the control of action requires predictive mechanisms (i.e. internal forward models) which in turn require a preselection of relevant sensory information.

The insights emerging from work on perception and action, at both cognitive and neural levels are of major importance for the understanding of human Mind. As a matter of fact, some philosophers of Mind, such as Goldman (1992) or Currie and Ravenscroft (1996) have proposed the idea that we predict the actions of others by simulating their decision-making processes. This view, called 'simulation theory' has generated considerable interest among cognitive scientists. Note that this view is close to the motor theories of perception, although not equivalent. Several authors have recently also argued, on the basis of experimental data, that our capacity to understand other people's behavior, to attribute intention of beliefs to others is rooted in a neural, most likely distributed, execution/observation mechanism (e.g. Gallese & Goldman 1998; Corballis 1998). Therefore, studies addressing perception of action and mental simulation of action (motor imagery), whether they are at the behavioral or at the neurophysiological levels, are of high relevance here and will thus be reviewed in this paper.

Behavioral evidence for perception of actions

There is plenty of evidence that the human visual system is finely attuned to human movements. A number of early studies, based on the point-light technique, have revealed that the kinematic pattern of a movement is sufficient for the perception of people in action (Johansson 1973). Using the same paradigm, Kozlowski and Cutting (1977) extended these findings by showing that observers can make very precise discriminations when watching point-light displays, such as the recognition of the gender of walkers. The fact that humans think in categories and that concepts influence the way of thinking and acting has lead Dittrich (1993) to investigate whether the ability to detect natural motions is in part determined by the content of independent categories of the information that physically characterize the event. In his study, locomotory (e.g. walking, going upstairs), instrumental (e.g. hammering, stirring) and social actions (e.g. greeting, boxing) were presented with the point-light display tech-

nique in a normal situation (light attached to joints), inter-joint (light attached between joints) and upside-down. Subjects' verbal responses and recognition times showed that locomotory actions were recognized much better and faster than social and instrumental actions. Furthermore, biological motions were recognized much better and faster when the light-spot displays were presented in the normal orientation rather than upside-down. Finally, the recognition rate was only slightly impaired under the inter-joint condition. The author argued that coding of dynamic phase relations and semantic coding take place at very early stages of the processing of biological motion.

It may thus be hypothesized that the recognition of biological motion is that perception and recognition processes are mediated by the implicit knowledge of production (motor) rules. Shiffrar and Freyd (1990, 1993) have performed a series of elegant studies by using the apparent motion paradigm (i.e. the impression of visual motion resulting from the sequential presentation of static objects in different spatial locations) with body pictures. They have challenged the fundamental important characteristic of classical apparent motion studies, namely that object identity does not influence the motion perceived since objects appear to move along the shortest or most direct path. Indeed, while static objects appear to move across 2-D and 3-D space, and over a wide range of stimulus presentation rates, apparent motion perception of human figures tends to respect the biomechanical constraints of normal human movement (see next section for neurophysiological evidence of this effect).

Difference in handedness may affect perception of action just as it affect motor performance. For instance, De'Sperati and Stucchi (1997) presented subjects with a rotating screwdriver on a computer screen and asked them whether it was screwing of unscrewing. Clear difference emerged in the response time between right- and left-handers.

Neurophysiological evidence for perception of actions

It is generally accepted that the monkey visual cortex is organized along two major anatomical and functional streams (Ungerleider & Mishkin 1982). The ventral stream projecting from V1 through areas V2 and V4 to the inferior temporal cortex and to the anterior section of superior temporal sulcus is considered to be primarily concerned with the recognition of objects. The dorsal stream, projecting from V1 through areas V2 and V3 to the middle temporal area (V5/MT) and thence to the superior temporal and parietal cortex is concerned with the perception of spatial information and with the visual guidance

of actions towards objects (Milner & Goodale 1995). The two pathways are not completely separate since a polysensory area in the superior temporal cortex (STPa) receives inputs both from the ventral and dorsal pathways (Boussaoud et al. 1990).

In the monkey, Perrett et al. (1989) have found that there are neurons in the superior temporal sulcus (STS) sensitive to the sight of static and dynamic information about the body. The majority of cells were selective for one perspective view and are considered to provide viewer-centered descriptions which can be used in guiding behavior. For some cells in the lower bank of STS (area Tea) the responses to body movements were related to the object or to the goal of the movements. The same group has also used the point light technique for presenting human biological motion to monkey while recording cells in STPa (Oram & Perrett 1997). One-third of the cells selective for the form and motion of walking body showed sensitivity to the moving light display.

Numerous electrophysiological recordings, performed by the Parma' s group, in the rostral part of monkey inferior premotor cortex (area F5) have indicated that there are neurons that discharge during execution of hand and mouth movements (e.g. Rizzolatti et al. 1988). Some years later, the same authors discovered that most of these neurons discharge not only when the monkey performed an action, but also when the monkey observed the experimenter making a similar action (e.g. Gallese et al. 1996). Neurons that exhibited such properties were therefore called 'mirror neurons'. Thus, the physiological results from STS and F5 neurons indicate that there are brain processes in monkeys which could support the understanding of other's actions, and that the sight of these actions could be matched to the motor commands in order to imitate or reproduce actions that monkey sees (Rizzolatti & Arbib 1998).

Another brain region involved in the visual analysis and control of action lies in the parietal cortex. Recent neurophysiological studies in alert monkeys have revealed that the parietal association cortex plays a crucial role in depth perception and in visually guided hand movement. Sakata et al. (1997) suggested that neural representation of 3-D objects with real physical dimensions and their egocentric positions seems to occur in the parietal association cortex. The major purpose of the 3-D representation is the visual guidance of goal directed action. Early indications are that parietal cells in area Tpa (homologous of the rostral part of inferior parietal cortex in human?) code the sight of action.

Thus from electrophysiological recordings in non-human primates, three key regions (namely in the superior temporal, parietal and ventral premotor

cortices) have been identified as specifically responsive to the visual perception of action and may have a role in action understanding.

In human, Fadiga et al. (1995) have demonstrated with magnetic transcranial stimulation an increase in excitability of the motor system (MEP) during the perception of actions performed by another individual. This enhancement of MEP is selective since it occurred in the muscles that the subjects would use for producing the action observed. Converging evidence has been reported by Cochin et al. (1998) in a study that used EEG cartography during the perception of different sorts video movies consisting of objects in movement, animal in motion, gymnastic movements executed by a person and still shots. Significant decreases in the alpha 1, beta 1 and beta 2 power values of the EEG over the centro-parietal regions, in both hemispheres, were shown during the perception of human motion sequences. Their results suggest the specific participation of the sensorimotor cortex during the observation of human motion. Magnetoencephalographic recordings have come to the same conclusion (Hari et al. 1998).

There have been a number of brain imaging studies that have addressed the question of the neural substrate involved in the perception of movement and action. Howard et al. (1996) have performed a functional magnetic resonance imaging (fMRI) study in which they contrasted different motion displays including biological motion. They found that each of the motion stimuli (i.e. coherent, optic flow and biological motion) activated specific parts of the V5/MT complex. The biological motion stimulus produced an area of activation that overlapped with the other two and was located along the superior border of V5. Both optical flow and biological motion stimuli produced small areas of activation within both dorsal and ventral V3, but neither produced any differential activation of area V1. This study demonstrates that there is a specialization for visual motion within the V5 complex and may help to account for some odd clinical observations of patients unable to perceive objects in motion while they can correctly identify biological motion (e.g. Vaina et al. 1990; Marcar et al. 1997). It should be emphasized that V5 responds to motion imagery, visual apparent motion (Goebel et al. 1998). An fMRI study designed by Grèzes et al. (2001) investigated the neurodynamic response to biological (non-rigid) versus rigid motion with point-light stimuli. They found gradient in activation in the occipito-temporal junction. The responses to rigid motion were localized posteriorly to those responses elicited by non-rigid motions. This study also reported that in addition to the posterior portion of superior temporal sulcus, the left intraparietal cortex is involved in the perception of non-rigid biological motions. It is also known that the generation of words denoting actions asso-

ciated with visually presented objects activates an area anterior to V5 which extends to the left posterior temporal gyrus (Martin et al. 1995).

Recently, on the basis of the findings by Shiffrar and Freyd (1990, 1993), we have adapted the apparent motion paradigm to present subjects with condition of possible and impossible bodily motion (Stevens et al. 2000). The V5 complex was found to be involved in both conditions. Moreover several other cortical areas were found to be selectively activated by each condition. When subjects perceived possible paths of apparent human movement the most significant bilateral rCBF increase was found in the primary sensorimotor cortex. Significant bilateral activations were also found in the superior parietal gyrus, superior frontal gyrus, superior temporal gyrus and in the left inferior parietal lobule. By contrast, perception of impossible apparent human motion lead to strong rCBF increases bilaterally in the medial orbitofrontal cortex. These results demonstrate that visual perception of apparent human movement selectively activates motor executive regions of the brain, just as explicitly perceived, imagined, and executed actions do. The selective activation of the orbital frontal cortex during the perception of impossible human movements likely reflects subjects' detection of deviations from normal motor action paths. Significant activation in this region has been found during perception of violations in visual tasks outcomes (Nobre et al. 1999) and when subjects plan for but then subsequently inhibit themselves from completing physical action (Krams et al. 1998).

Additional evidence about the neurophysiological substrate underlying the perception of action in human comes from several functional neuroimaging experiments. Rizzolatti et al. (1996) scanned subjects under three experimental conditions: observation of an actor grasping common physical objects, grasping the same objects themselves and as a control, passive object observation. The results of subtracting object observation from observation of an actor grasping the object resulted in rCBF activations in the middle temporal gyrus including that of adjacent superior temporal sulcus, in the caudal part of the inferior frontal gyrus, as well as activations in the precuneus and in the medial frontal gyrus. All activations were located in the left hemisphere. Their data show that there is a good correspondence between humans and monkeys in the cortical areas devoted to hand action recognition. The activation in the left temporal lobe would correspond to the STS in monkey (Perrett et al. 1989, 1990; Oram et al. 1994) and the activation in the inferior frontal gyrus might be similar to F5 (e.g. Gallese et al. 1996). These results have been confirmed and extended by Decety et al. (1997) in a PET experiment that contrasted the visual perception of meaningful and meaningless pantomimes. Subjects were

instructed to watch the actions with two aims: either to recognize or to imitate them later. It was found that the meaning of the action, irrespective of the strategy used during observation lead to different patterns of brain activity and clear left/right asymmetries. Meaningful actions strongly engaged the left hemisphere in frontal and temporal regions while meaningless actions involved mainly the right occipito-parietal pathway. Observing with the intent to recognize activated memory encoding structures. By contrast observation with the intent to imitate was associated with activation in the regions involved in the planning and in the generation of actions. Thus, the pattern of brain activation during observation of actions is dependent, both, on the nature of the required executive processing and the type of the extrinsic properties of the action presented. Another PET study, conducted by Grèzes et al. (1998) used a similar paradigm. Perception of meaningful and of meaningless hand actions without any purpose was contrasted with the perception of the same kind of stimuli with the goal to imitate them latter. A condition with perception of stationary hands served as a baseline level. Perception of meaningful actions and meaningless actions without any aim was associated with activation of a common set of cortical regions. In both hemispheres, the occipito-temporal junction and the superior occipital gyrus were involved. In the left hemisphere, the middle temporal gyrus and the inferior parietal lobe were found to be activated. The precentral gyrus, within the somatotopic representation of the hand area was activated in the left hemisphere. In addition to this common network, meaningful and meaningless movements engaged specific networks, respectively: meaningful actions were associated with activations mainly located in the left hemisphere in the inferior frontal gyrus and the fusiform gyrus, whereas meaningless actions involved the dorsal pathway bilaterally and the right cerebellum. In contrast, meaningful and meaningless actions shared almost the same network when the aim of the perception was to imitate. Activations were located in the right cerebellum and bilaterally in the dorsal pathway reaching the premotor cortex. Additional bilateral activations were located in the SMA and in the orbitofrontal cortex during observation of meaningful actions. Thus, when perception has no goal, the pattern of brain activation is dependent on the nature of the movements presented. But when perception has a goal, namely to imitate, the subject's strategy has a top-down effect on the information processing which seems to give priority to the dorsal pathway involved in perception for action (Goodale 1997).

Behavioral evidence for mental simulation of actions

The involvement of common neural ressources in motor imagery and motor behavior is supported by a growing body of evidence since the seminal paper by Decety and Ingvar (1990). In this paper we suggested that mental simulation of action requires the construction of a dynamic motor representation in working memory which makes use of spatial and kinesthetic components retrieved from long-term memory, as well as the activation of serial plans of action.

Several experiments using the mental chronometry paradigm clearly support the hypothesis that mental simulation of action is assigned to the same motor representation system as preparation and execution. Notably Parsons et al. (1987) have shown that, when a photograph of a hand at a given orientation was presented at a given orientation and the subjects required to judge whether it is a right or a left hand, the time taken to respond closely mimic the actual time of the real movement. In subjects requested either to imaging or actually walk towards targets placed at different distances, Decety et al. (1989) reported that the duration of simulation was similar and related to the distance covered. Another evidence for the implementation of motor production rules during mental simulation has been demonstrated with the use of Fitts law paradigm (i.e. the tradeoff between speed and accuracy). For instance, Decety and Jeannerod (1996) designed an experiment to verify whether Fitts law hold in mental simulation. Subjects, immersed in a virtual reality system, were instructed to walk mentally along a path and through gates of different widths positioned at different distances. In accordance with Fitts' law, the time needed to imagine walking through a given gate was found to be affected both by its relative distance and by its relative width, namely the duration of the simulation increased linearly as a function of task difficulty. Different performance in handedness also affect mental simulation as it does it motor performance (e.g. De'Sperati & Stucchi 1997). In a recent experiment, Johnson (1998) exploited the well-known contralateral organization of the visual and motor systems with a divided-visual-field task that required subjects to determine whether they would use an underhand or overhand grip if the were to reach for a dowel-shaped object, briefly presented to either visual fields. Although no actual reaching movements were performed, a significant advantage in grip-selection time was found when the information was presented to the cerebral hemisphere contralateral to the designated response hand.

There is converging evidence from studies of neurological patients which demonstrates that deficit in motor performance may be reflected in motor imagery. Interestingly, Dominey et al. (1995) examined hemi-Parkinson patients

in both visual and motor imagery tasks involving either side of the body. They reported a selective deficit in motor imagery on the affected side but not in visual imagery that closely matched the deficit in actual motor performance. A correlation between actual and mental movement times was found in a patient with motor cortex damage in the right hemisphere (Sirigu et al. 1995). The left arm was slower in executing motor tasks with the fingers and elbow, but not with the shoulder. The same difference was observed for mentally simulated movements. Sirigu et al. (1996) also reported that patients with lesions restricted to the parietal cortex were found to be selectively impaired at predicting, through mental imagery tasks, the time necessary to perform finger movements, in comparison to normal subjects and to the previously described patient with damage to the primary motor area. A similar observation has been reported in a single case study of a patient with severe ideomotor apraxia who was selectively impaired in motor imagery while his capacity in visual imagery of objects was spared (Ochipa et al. 1997).

Neurophysiological evidence for mental simulation of actions

Studies of cerebral metabolic activity have shown overlapping patterns of rCBF variations during motor imagery, motor preparation and actual motor performance. In a group of normal subjects requested to imagine grasping objects relative to the visual inspection of the same objects the prefrontal cortex, the anterior cingulate, the premotor cortex, the inferior parietal lobule, the cerebellum, the ventrolateral thalamus and the caudate nucleus were found to be activated mainly in the left hemisphere (Decety et al. 1994). Lang et al. (1994) reported bilateral activations in the SMA, in the precentral gyrus, and in the anterior cingulate gyrus during the simulation of saccadic eye movements. Another PET activation study during an internally guided motor imagery of a joy-stick movement reported activations in the medial and lateral premotor areas, including the SMA as well as the superior and inferior parietal areas bilaterally. (Stephan et al. 1995). Until recently, there was agreement that the early stages of the process of action generation, such as planning or programming, involved the premotor cortex and the SMA, while MI was considered to be responsible for proper execution only. This view now has been challenged by the use of functional magnetic resonance imaging (fMRI), whose temporal and spatial resolution is much better than PET. For example, Roth et al. (1996) measured hemodynamic changes with fMRI in normal right-handed subjects during actual and mental execution of a finger-to-thumb opposition task per-

formed with either the right or the left hand. The data show no significant differences between the two hands with either execution or simulation. A significant involvement of contralateral motor cortex (30 % of the activity found during execution) was detected. The premotor cortex and the rostral part of the posterior SMA were activated bilaterally during motor imagery.

Few neuroimaging studies have examined the neural correlates of motor preparation in humans. Decety et al. (1992) instructed subjects to prepare for pointing towards visual targets previously displayed on a screen and found bilateral activations in the prefrontal cortex, SMA, left parietal lobe and supramarginal gyrus, the ventrolateral thalamus and the cerebellum. Recently, Krams et al. (1998) measured movement set-related changes when subjects were asked to copy hand movements during four conditions: prepare and execute, immediate execution, preparation only and a baseline condition. They demonstrated that regions involved in motor preparation were a subset of the areas involved in the preparation and execution condition. However the finding that the ventrolateral prefrontal cortex and the anterior cingulate were significantly more active during preparation than in a coupled preparation/execution condition and was interpreted as inhibitory mechanism.

The activation of representations for actions by motor imagery is not limited to increased metabolism in brain areas. It also includes autonomic mechanisms, which represent an interesting case since they are normally not submitted to voluntary control. Indeed several papers have reported cardiorespiratory changes during motor imagery, namely increases in heart rate, respiratory rate, ventilation and blood pressure (e.g. Decety et al. 1993; Wang & Morgan 1992). The same phenomenon has been observed during motor preparation (see Requin et al. 1991).

Discussion

An action may be described as the outcome of several information processing stages: intention, planning, preparation and eventually execution. It is widely accepted that an internal model underlies these mechanisms. Several cognitive models have been put forward to account for the capacity to generate an action but also to understand, predict actions performed by others and even distinguish our own actions from events caused by external agents (e.g. Frith 1992; Jeannerod 1997; Blakemore & Decety 2001). This is not restricted to motor actions but can be generalized to the expression and recognition of emotions. Indeed there is considerable evidence we tend to imitate (not consciously) fa-

cial expressions to which we are exposed, via feedback mechanisms we realize that our own imitated facial expression is associated with an emotion, and then attribute this emotion to the person confronting us (e.g. Wallbott 1991). An attractive hypothesis is that a similar neural system operates for action production and action recognition (e.g. Rizzolatti & Arbib 1998; Gallese & Goldman 1999; Blakemore & Decety 2001). Such a system is likely to be distributed among different brain areas. Indeed, a common network seems to be consistently recruited during several action modalities, namely the inferior parietal lobule, the prefrontal and premotor cortex and perhaps the primary motor cortex, although there is no strict overlap between each modality (Grèzes & Decety 2001).

Yet the respective role of each of these components that are making this system work is not clearly understood in human. Even if it is true that we do gain clues from the detailed comparison of the human brain with that of the macaque monkey, those homologies have to be carefully employed since less is known in the chimpanzee brain, our nearest relative according to genetic and comparative psychology studies (see Passingham 1998). For instance, if area F5, in which mirror neurons are recorded in monkey, has evolved to Broca's area in man as suggested by MacNeilage (1998), then one should hypothesize that lesion of the left inferior frontal gyrus would lead to deficits in gesture recognition and in gesture imitation. Several clinical observations since the time of Jackson (1878) have indeed reported oral apraxia to be associated with Broca's aphasia but not with Wernicke's aphasia (e.g. De Renzi et al. 1966). Matteer and Kimura (1977) tested nonfluent aphasics as well as fluent aphasics in the imitation of nonverbal oral movements tasks. They found that, while the former were impaired in imitating single oral movements, the later showed impairments in the imitation of complex oral movements. An interesting question is whether such impairment is specific to oral movements *per se* or can be extended to hand gestures. Very few systematic studies addressed this issue. Bell (1994) has presented a pantomime recognition test (involving a combination of limb, hand and facial actions) to a group of aphasic patients (regardless of various forms of aphasia, i.e. Broca's, conduction, Wernicke's, anomic, global, nonfluent) and compared their results to a group of control subjects. Patients performed significantly poorer than the controls. However, when one looks carefully at the data reported in the article, only 10 patients out of 23 scored incorrectly on the recognition test. We have recently examined the ability to recognize and to imitate meaningful and meaningless pantomimes in a group of neurological patients sorted by the location of the vascular lesion, as opposed to a group of control subjects (Genouilhac 1998). While all patients' per-

formance in the recognition of meaningless and meaningful actions tasks was comparable to that of normals regardless of the laterality of the lesion, only one case (left paramesial protuberance lesion) was impaired. We have since examined three Broca's patients in various systematic recognition and imitation tasks involving object-directed movements, meaningful and meaningless pantomimes. None of the patients exhibited impairment in any of these tasks. Patients with left hemisphere infarct involving Broca's area and surrounding regions were not impaired in ordering groups of words to form a logical sequence of action (Sirigu et al. 1998). The test they used consisted of a set of cards with an action written on each one. Hence the patients had first to recognize the action before ordering it into a sequence. Thus there is clearly a need for more studies of patients to assess the role of the inferior frontal gyrus in action recognition and so far there is no serious evidence that this region is as crucial as the mirror neurons hypothesis assumes from monkey work.

The middle temporal gyrus (Brodmann area 21) is another region of interest for the perception of action. Here again, the homologies between monkey and human should be taken with caution. While in macaque areas 21 and 22 act as visual association cortex, in human, area 21 is engaged in language tasks (e.g. Demonet et al. 1992; Martin et al. 1995) and in the perception of hand and body movements in its posterior portion (e.g. Grafton et al. 1996; Decety et al. 1997; Grèzes et al. 1998). This region is distinct from V5 and is found predominantly activated in the right hemisphere. Puce et al. (1998) have also reported that a region in the posterior part of the temporal lobe, in the STS, was activated when subjects view a face in which the eye or mouth are moving. These activations were not attributable to movement *per se* but to the fact that the movement was biological. Thus, there is no doubt that an area preferentially responsive to biological movements is present in human STS.

Conclusion

We know from evolutionary psychology that there is continuity between non-human and human primate. It is thus plausible that our ability to recognize and attribute intentions to ourselves as well as to others has evolved from the capacity to perceive and identify movements performed by other living organisms. This system exploits the same motor representations for both reading others' actions (in a wide sense, e.g. gaze direction, emotion) and for reproducing them.

It is acknowledged that newborns are wired for perceptual-motor coupling as it is attested by the studies on immediate imitation (Meltzoff & Moore 1999). Imitative interaction sequences play a constitutive role in intersubjectivity as well as in the learning of social rules (see Nadel & Butterworth 1999). The prefrontal cortex which appears to exert its functions mostly through inhibition is not fully mature immediately after birth. While cytoarchitecture reaches full development before birth in human, the myelination of prefrontal connective fibers extends long after birth, until adolescence (Fuster 1997). This lack of inhibition, or mild inhibition at the beginning of childhood confers developmental benefits through imitation. Then, inhibitory mechanisms progressively develop, in parallel to cognitive abilities for which inhibition is a requisite. Substantial evidence for the inhibitory role of the lower half of the prefrontal cortex has been reported in brain damaged individuals (Lhermitte et al. 1996) who exhibit unsuppressed imitation or utilization behavior. Interesting by the orbitofrontal region was consistently found to be activated when healthy subjects watched actions for later imitation (Decety et al. 1997; Grèzes et al. 1998) and not when subjects observed for later recognition. Furthermore this activation was detected only for those actions that were in the motor repertoire of the subjects, e.g. meaningful actions, but not for meaningless actions. In addition it was recently demonstrated that when meaningless actions were learned by subjects a few days before the PET exam and thus became familiar, the orbitofrontal cortex was then engaged during observation with the intent to imitate. It was not the case for unfamiliar meaningless actions although the difference between the two sets of movements was very tiny (Grèzes & Decety 1999).

The lines of evidence presented in this chapter are largely compatible with the idea that neural and cognitive representations involved in actions generation are also recruited by observation and by mental simulation. Thus perception and action are not separate phenomena. Rather there is a common neural substrate that directly translates sensory experience into action or schemas of actions. Whether the overlap between the cortical regions is true not only at a macro-anatomical but chiefly at a micro-anatomical level remains to be elucidated (this will be the case in the near future).

Finally if one accepts the idea of shared representations for recognition, imitation, or self production then how do we distinguish our own actions from the actions from the environment? Several researchers have proposed that knowledge of our intentions or motor commands is used to distinguish the sensory consequences of our own actions from externally-produced sensory stimuli (Frith 1992; Wolpert et al. 1995). Frith has argued for a central

monitoring system, which can be considered as a cognitive model similar to the forward model but with an extension of the functional properties of the efference copy concept into to covert actions such as thinking. This model captures the forward or causal relationship between actions, as signaled by the efference copy and the predicted sensory outcome. By comparing this prediction with the actual sensory feedback it is possible to distinguish the sensory consequences of our movements from sensory signals due to changes in the outside world. Different cortical areas for expected stimuli and sensory consequences of self-generated actions have been discovered by Blakemore et al. (1998). They have observed an interaction between the predictability of the stimuli and self-generated actions in the medial posterior cingulate cortex, the left insula, the dorsomedial thalamus and the right inferior temporal cortex. In an fMRI study, Blakemore et al. (1999) examined neural responses when subjects experienced a tactile stimulus that was either self-produced or externally-produced. More activity was detected in somatosensory cortex when the stimulus was externally-produced and a reduction of activity was seen for self-produced action. There was less activity in the cerebellum when a movement generated a tactile stimulus than with a movement that did not. This difference suggests that the cerebellum is involved in predicting the specific sensory consequences of movements, providing the signal that is used to cancel the sensory response to self-generated stimulation. Such selective activity for externally generated signal are also evident at the single cell level.

Another mechanism for discriminating self-produced action and observed action is based on the partial overlap between cortical networks for different modalities of action representation has been proposed by Jeannerod (1999). Activation of those areas which overlap during a self-produced and an observed action would be interpreted as an observed action. By contrast, activation of nonoverlapping areas would be interpreted as a self-produced action. However, in the light of the neuroimaging data reviewed here, I would suggest that the difference between the cortical areas engaged during action recognition and during action generation is, for many of them, more a difference of degree of activation (number of neurons, discharge frequency) rather of nature. In addition, neuroimaging studies have very recently pointed out a specific role played by the inferior parietal cortex in agency (i.e. the awareness of being the cause of the action we produce). The left inferior parietal cortex is more involved when the other is the cause of the action performed by the self, whereas the right homologous region is more activated when the self is the source of the action while watching it performed by the other, as demonstrated by means of mutual imitation paradigm (Decety et al. 2002; Chaminade & Decety, in press)

as well as a study of perspective-taking (Ruby & Decety 2001). Finally, the neural structures involved in the inhibition of action play a fundamental role in this model.

References

Bell, B. D. (1994). Pantomime recognition impairment in aphasia: an analysis of error types. *Brain and Language, 47*, 269–278.

Blakemore, S. J., Rees, G., & Frith, C. D. (1998). How do we predict the consequences of our actions? A functional imaging study. *Neuropsychologia, 36*, 521–529.

Blakemore, S. J., Wolpert, D. L., & Frith, C. D. (1999). Central cancellation of self-produced tickle sensation. *Nature Neuroscience, 1:7*, 635–640.

Blakemore, S. & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*, 561–567.

Boussaoud, D., Ungerleider, L. G., & Desimone, R. (1990). Pathways for motion analysis: cortical connections of the medial superior temporal and fundus of the superior temporal visual areas in the macaque. *Journal of Comparative Neurology, 296*, 462–495.

Chaminade, T. & Decety, J. (2002). Leader or follower? involvement of the inferior parietal lobule in agency. *Neuroreport*, in press.

Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical Neurophysiology, 107*, 287–295.

Corballis, M. C. (1998). Evolution of the human mind. In M. Sabourin, F. Craik, & M. Robert (Eds.), *Advances in Psychological Science, Volume 2*: *Biological and cognitive aspects* (pp. 31–62). Hove, UK: Psychology Press.

Currie, G. & Ravenscroft, I. (1997). Mental simulation and motor imagery. *Philosophy of Science, 64*, 161–180.

Davidson D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

De'Sperati, C. & Stucchi, N. (1997). Recognizing the motion of graspable object is guided by handedness. *Neuroreport, 8*, 2761–2765.

Decety, J. & Jeannerod, M. (1996). Mentally simulated movements in virtual reality: does Fitts's law hold in motor imagery? *Behavioural Brain Research, 72*, 127–134.

Decety, J. & Ingvar, D. H. (1990). Brain structures participating in mental simulation of motor behavior: A neuropsychological interpretation. *Acta Psychologica, 73*, 13–34.

Decety, J., Perani, D., Jeannerod, M., Bettinardi, V., Tadary, B., Woods, R., Mazziotta, J. C., & Fazio, F. (1994). Mapping motor representations with positron emission tomography. *Nature, 371*, 600–602.

Decety, J. (1996). Do imagined and executed actions share the same neural substrate? *Cognitive Brain Research, 3*, 87–93.

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*, 1763–1777.

Decety, J., Prablanc, C., & Jeannerod, M. (1989). The timing of mentally represented actions. *Behavioral Brain Research, 34*, 35–42.

Decety, J., Kawashima, R., Gulyas, B., & Roland, P. (1992). Preparation for reaching: A PET study of the participating structures in the human brain. *NeuroReport, 3*, 761–764.

Decety, J., Jeannerod, M., Durozard, D., & Baverel, G. (1993). Central activation of autonomic effectors during mental simulation of motor actions in man. *Journal of Physiology, 461*, 549–563.

Decety, J., Chaminade, T., Grèzes, J., & Meltzoff, A. (2002). A PET exploration of the neural mechanisms involved in reciprocal imitation. *Neuroimage, 15*, 265–272.

Demonet, J. F., Chollet, J. F., Ramsay, S., Cardebat, D., Nespoulous, J. L., & Wise, R. (1992). The anatomy of phonological processing in normal subjects. *Brain, 115*, 1753–1768.

De Renzi, E., Pieczuro, A., & Vignolo, L. A. (1966). Oral apraxia and aphasia. *Cortex, 2*, 50–73.

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception, 22*, 15–22.

Dominey, P., Decety, J., Broussolle, E., Chazot, G., & Jeannerod, M. (1995). Motor imagery of a lateralized sequential task is asymmetrically slowed in hemi-Parkinson patients. *Neuropsychologia, 33*, 727–741.

Fadiga, L., Fogasi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. LEA.

Fuster, J. M. (1997). *The prefrontal cortex*. Lippincott – Raven, Philadelphia.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 593–609.

Gallese, V. & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*, 493–501.

Genouilhac, V. (1998). La perception des actions chez des patients cérébrolésés. Master Thesis. Institute of Psychology, University Lumière, Lyon 2.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience, 10*, 1563–1573.

Goldman, A. I. (1992). In defense of the simulation theory. *Mind and Language, 7*, 104–119.

Goodale, M. A. (1997). Visual routes to perception and action in the cerebral cortex. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology*, Vol. 11 (pp. 91–109). Elsevier.

Grafton, S. T. (1996). Localization of grasp representations in humans by positron emission tomography. *Experimental Brain Research, 112*, 103–111.

Grèzes, J., Costes, N., & Decety, J. (1998). Top-down effect of the strategy on the perception of biological motion: A PET investigation. *Cognitive Neuropsychology, 15*, 553–582.

Grèzes, J., Costes, N., & Decety, J. (1999). The effect of learning on the neural networks engaged by the perception of meaningless actions. *Brain, 122*, 1875–1888.

Grèzes, J., Fonlupt, P., Bertenthal, B., Delon-Martin, C., Segebarth, C., & Decety, J. (2001). Does perception of biological motion rely on specific brain regions? *Neuroimage, 13*, 775–785.

Grèzes, J. & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation and verb generation of actions: A meta-analysis. *Human Brain Mapping, 12*, 1–19.

Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: A neuromagnetic study. *Proceedings Natl. Acad. Sci. USA, 95*, 15061–15065.

Howard, R. J., Brammer, M., Wright, I., Woodruff, P. W., Bullmore, E. T., & Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. *Current Biology, 6*, 1015–1019.

Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Oxford: Blackwell.

Jeannerod, M. (1999). To act or not to act: perspectives on the representation of actions. *The Quaterly Journal of Experimental Psychology, 52A*, 1–29.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14*, 201–211.

Johnson, S. H. (1998). Cerebral organization of motor imagery: contralateral control of grip selection in mentally represented prehension. *Psychological Science, 9*, 219–222.

Jordan, M. I. (1994). Computational motor control. In M. Gazzaniga (Ed.), *Cognitive Neuroscience* (pp. 597–699). MIT Press.

Kozlowski, L. T. & Cutting, J. E. (1977). Recognizing the sex of a walker from point-lights display. *Perception and Psychophysics, 21*, 575–580.

Krams, M., Rushworth, M. F. S., Deiber, M.-P., Frackowiak, R. S. J., & Passingham, R. E. (1998). The preparation, execution and suppression of copied movements in the human brain. *Experimental Brain Research, 120*, 386–398.

Lang, W., Petit, L., Höllinger, P., Pietrzyk, U., Tzourio, N., Mazoyer, B., & Berthoz, A. (1994). A positron emission tomography study of oculomotor imagery. *NeuroReport, 5*, 921–924.

Lhermitte, F., Pillon, B., & Serdaru, M. (1986). Human autonomy and the frontal lobes. Part I: Imitation and utilization behavior: A neuropsychological study of 75 patients. *Annals of Neurology, 19*, 326–334.

Marcar, V. L., Zihl, J., & Cowey, A. (1997). Comparing the visual deficits of a motion blind patient with the visual deficits of monkeys with area MT removed. *Neuropsychologia, 35*, 1459–1465.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences, 21*, 499–546.

Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., & Ungerleider, L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science, 270*, 102–105.

Meltzoff, A. N. & Moore, M. K. (1999). Persons and representation: Why infant imitation is important for theories of human development. In J. Nadel & G. Butterworth (Eds.), *Imitation in infancy* (pp. 9–35). Cambridge University Press.

Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action*. Oxford: Oxford University Press.

Nadel, J. & Butterworth, G. (Eds.). (1999). *Imitation in infancy*. Cambridge University Press.

Nobre, A. C., Coull, J. T., Frith, C. D., & Mesulam, M. M. (1999). Orbitofrontal cortex is activated during breaches of expectation in tasks of visual attention. *Nature Neuroscience, 2*, 11–12.

Ochipa, C., Rapesak, S. Z., Maher, L. M., Rothi, L. J. G., Bowers, D., & Heilman, K. M. (1997). Selective deficit of praxis imagery in ideomotor apraxia. *Neurology, 49*, 474–480.

Oram, M. W. & Perrett, D. I. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to biological motion stimuli. *Journal of Cognitive Neuroscience, 6*, 99–116.

Parsons, L. M. (1987). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology, 19*, 178–241.

Passingham, R. E. (1998). The specializations of the human neocortex. In A. D. Milner (Ed.), *Comparative Neuropsychology* (pp. 271–298). Oxford: Oxford University Press.

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., Chitty, A. J., Hietanen, J. K., & Ortega, J. E. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology, 146*, 87–114.

Perrett, D. I. (1990). Understanding the visual appearance and consequence of actions. In M. A. Goodale (Ed.), *Vision and Action* (pp. 163–180). Norwood, NJ: Ablex Publishing Corporation.

Prinz, W. (1992). Why don't we perceive our brain states? *European Journal of Cognitive Psychology, 4*, 1–20.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129–154.

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of Neuroscience, 18*, 2188–2199.

Requin, J., Brener, J., & Ring, C. (1991). Preparation for action. In J. R. Jennings & M. G. H. Coles (Eds.), *Handbook of Cognitive Psychophysiology. Central and Autonomic Nervous System Approaches* (pp. 357–458). New York: John Willey & Sons.

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. *Experimental Brain Research, 71*, 491–507.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Perani, D., & Fazio, F. (1996). Localization of grasp representations in humans by PET: 1 Observation versus execution. *Experimental Brain Research, 111*, 246–252.

Rizzolatti, G. & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188–194.

Roth, M., Decety, J., Raybaudi, M., Massarelli, R., Delon, C., Segebarth, C., Gemignani, A., Decorps, M., & Jeannerod, M. (1996). Possible involvement of primary motor cortex in mentally simulated movement: A functional magnetic resonance imaging study. *NeuroReport, 7*, 1280–1284.

Ruby, P. & Decety, J. (2001). Effect of the subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience, 4*, 546–550.

Sakata, H., Taira, M., Kusunoki, M., Murata, A., & Tanaka, Y. (1997). The parietal association cortex on depth perception and visual control of hand action. *Trends in Neurosiences, 20*, 350–357.

Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.

Shiffrar, M. & Freyd, J. J. (1990). Apparent motion of the human body. *Psychological Science, 1*, 257–264.

Shiffrar, M. & Freyd, J. J. (1993). Timing and apparent motion path choice with human body photographs. *Psychological Science, 4*, 379–384.

Sirigu, A., Cohen, L., Duhamel, J. R., Pillon, B., Dubois, B., Agid, Y., & Pierrot-Deseilligny, C. (1995). Congruent unilateral impairments for real and imagined hand movements. *Neuroreport, 6*, 997–1001.

Sirigu, A, Duhamel, J. R., Cohen, L., Pillon, B., Dubois, B., & Agid, Y. (1996). The mental representation of hand movements after parietal cortex damage. *Science, 273*, 1564–1568.

Sirigu, A., Cohen, L., Zalla, T., Pradat-Diehl, P., Van Eechout, P., Grafman, J., & Agid, Y. (1998). Distinct frontal regions for processing sentence syntax and story grammar. *Cortex, 34*, 771–778.

Sperry, R. W. (1952). Neurology and the mind-brain problem. *American Scientist, 40*, 291–312.

Stephan, K. M., Fink, G. R., Passingham, R. E., Silbersweig, D., Ceballos-Baumann, A. O., Frith, C. D., & Frackowiak, R. S. J. (1995). Functional anatomy of the mental representation of upper extremity movements in healthy subjects. *Journal of Neurophysiology, 73*, 373–386.

Stevens, J. A., Fonlupt, P., Shiffrar, M., & Decety, J. (2000). New aspects of motion perception: selective neural encoding of apparent human movements. *NeuroReport, 11*, 109–115.

Ungerleider, L. G. & Mishkin, M. (1982). Two visual systems in Analysis of visual behavior. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), (pp. 549–586). Cambridge, MA: MIT Press.

Vaina, L. M. LeMay, M., Bienfang, D. C., Choi, A. Y., & Nakayama, K. (1990). Intact biological motion and structure from motion perception in a patient with impaired motion mechanisms: A case study. *Visual Neuroscience, 5*, 353–369.

Wallbott, H. G. (1991). Recognition of emotion from facial expression via imitation? Some indirect evidence for an old theory. *British Journal of Social Psychology, 30*, 207–219.

Wang, Y. & Morgan, W. P. (1992). The effects of imagery perspectives on the physiological responses to imagined exercise. *Behavioural Brain Research, 52*, 167–174.

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

Reply to Jean Decety

Perceiving actions and understanding agency

Christoph Hoerl
University of Warwick

1. Let me start this commentary by quoting a passage from a paper by Andrew Meltzoff and Keith Moore:

> The human act may be the earliest, most aboriginal parsing of the world into things that bear on the self and those that do not. [...] When a human act is shown to a newborn baby, the act may provide its first 'aha' experience. 'Lo! Something interpretable. That (seen) event is like this (felt) event.' The fact that infants can recreate the act allows them to give it special meaning. (Meltzoff & Moore 1995: 54)

One way of reading this passage is in terms of the claim that our ability to imitate others' actions holds the key to answering the question as to how we come to understand what it is for others to be like us, and for us to be like them. On closer inspection, however, the precise way in which imitation is supposed to come in here is quite unclear.

On the face of it, it might seem as though the 'aha' experience Meltzoff and Moore talk about arises from the ability to *interpret* what we see others doing as being just like what we can feel ourselves doing. Thus understood, it would seem that we first have to have the required 'aha' experience before we can imitate others. Imitation, on this view, would at best be something like the most fundamental way of putting to use our understanding of others as being like us, but it could hardly be said to explain what makes such an understanding possible. Alternatively, we might think that it is simply a brute fact that we *feel* just the same when we watch someone doing something as when we do it ourselves (or, as we would feel if we did it ourselves), and that this can explain the 'aha' experience. On this view, imitation might be seen as a way of greeting

a perceived action as familiar by recreating it. Yet, if this is right, what makes the two actions different, what makes the perceived action yours and the imitation of it mine, is in danger of dropping out of the picture.

Meltzoff and Moore are clearly aware of these problems. What I am interested in here is the way they try to solve them. For them, the problem is how our own actions and those of others are represented. Very roughly, the point is as follows. In some sense, our own and others' actions must be represented the same way, or else it would remain mysterious how we ever come to understand what makes us and others the same. Yet, in another sense, our own and others' actions must be represented differently, or else it would remain mysterious how we ever come to grasp what makes some actions ours and others those of other people.

What Meltzoff and Moore suggest, in effect, is that we can reconcile these two points by drawing a distinction between two levels at which questions about representations can be addressed.[1] When Meltzoff and Moore talk about the sense in which we can be said to represent our own and others' actions in the same way, they say that they are represented in the same mental code. What this means, I take it, is that imitation does not really require *us* to forge a link between what we see others doing and what we can do ourselves. The correspondence between the two is written into the causal structure of our brains.[2] By contrast, when Meltzoff and Moore talk about the sense in which we can be said to represent our own and others' actions differently, they say that it is up to us whether we imitate someone or not, that we can correct our imitations, or recognize being imitated by others. Here, it *is* us who are described as doing certain things, not our brains, and it is with respect to this aspect that our understanding of others as being like us is ultimately a matter of what we can do.

Setting things up in this way suggests a certain division of labour in the empirical study of imitation. On the one hand, we would like neuroscience to give us some idea of what it means to say that actions we see others perform and actions we perform ourselves are represented in the same mental code, by looking at what happens in the brain when we perceive and imitate the actions of others. On the other hand, we would like developmental psychology to provide us with some story as to how children's ability to understand others as being like them, and them as being like others, develops, by charting the growing sophistication of imitative behaviour in children.[3]

2. Jean Decety presents a broad array of behavioural and neurophysiological data which support the idea of a fundamental connection, on the level

of the brain, between perceiving actions and carrying out actions oneself. In what follows, I wish to concentrate on two questions: To what extent do these findings support the idea of a common mental code for perceived actions and actions carried out by the subject herself; and, to the extent that they do so, can they also help us make sense of the idea that there may be different levels of sophistication or different steps in the development of our understanding of agency?

The first thing to note, I think, is that the claim that perceived actions and performed actions are represented in the same mental code is more specific than the claim that there is a link, on the level of brain mechanism, between perceiving an action and performing an action.[4] I am thinking here, for instance, of experiments involving so-called Johansson figures (see Johansson 1973). In those experiments, participants are shown a display in which all that can been seen are a number of point-lights which move in a certain way with respect to each other. The display is actually achieved by filming a person whose body is covered in black cloth against a black background. The point-lights are usually fixed to the person's joints, and their relative movement results from the movements the person carries out. The finding is that participants can make very precise discriminations just from watching this kind of point-light display. The speed at which they can do so depends, to some extent, on the nature of the activity to be recognized, but, basically, it turns out that we judge that there is a person doing certain things on the basis of very limited information about certain patterns of movement.

On the basis of these observations Decety hypothesises that, in the recognition of biological motion, perception and recognition processes are mediated by the implicit knowledge of certain production rules. The question I wish to raise is whether these experiments necessarily show that these rules are the same that underlie production of our own behaviour, as the idea of a common mental code for perception and action (at least according to Meltzoff's & Moore's construal) would seem to demand.

Other animals (including other human animals) are the kinds of things we can often only see bits of. Sometimes all we can see is a particular pattern of motion, without actually seeing the animal itself, as when a snake is slithering through high grass. Yet, what we are dealing with, whether it is friend or foe or food, and what they are up to, can be extremely important for us and our survival. It is therefore not surprising if we find that mechanisms have evolved that are keyed to the recognition of biological motion, and are linked directly with action. But talking about a direct link, here, is not the same as talking about a common mental code in which both perceived action and appropriate

response are represented. For one thing, the appropriate response is rarely of the same kind as the perceived action. Moreover, if what is perceived is the action of an animal of a different species, there may not even *be* a common motor pattern shared by the action we perceive and any actions we could carry out.[5]

The general point here is that experiments involving Johansson figures may show that we have some kind of implicit knowledge of the principles underlying perceived behaviour. Furthermore, this kind of knowledge may indeed have a crucial role to play in various forms of psychological understanding. Here we might think, for instance, of the kind of practical grasp of the significance of others' behaviour that Niko Frijda describes when he says that "understanding a given reaction as anger may consist of readyng oneself for attack" (Frijda 1969: 169).[6] However, while this kind of practical grasp requires *some* causal link between perception and action, it does not require that perceived actions and self-initiated actions are represented in the same mental code.

3. Let me now turn to some of the other findings mentioned by Decety, which appear to provide much firmer footing for the claim that perceived actions and self-initiated actions are represented in the same mental code (e.g., Fadiga et al. 1995; Gallese et al. 1996). What these findings suggest, in effect, is that the actual brain mechanisms involved in perceiving an action and carrying out that action oneself overlap (at least under certain conditions).

Why should such findings be important? I think one reason why research in this area is interesting is because of what it might be able to show us about the difference between the kind of practical grasp of others' activity described in the previous section, on the one hand, and imitation, on the other. Imitation involves more than just responding to the other's action in some appropriate way, where responding in the appropriate way just happens to consist in reproducing the same action.[7] And what I wish to suggest is that the idea of a common mental code can make intelligible a sense in which it does so. In short, I wish to argue that there is a connection between the idea of a common mental code involved in perceiving and performing the same action, as something that can be studied empirically, and our common sense understanding of what imitation is.

I think there are important parallels between this issue and an issue which stands at the heart of so-called 'architecturalist' theories of the mind (cf. Davies 1991). One of the questions that these theories try to address may be illustrated as follows. In what way must Paris be represented in my belief 'Paris is the capital of France' and my belief 'The Louvre is in Paris', if I am to be able to draw the conclusion that the Louvre is in the capital of France? What we need to

explain, in short, is the sense in which my arriving at the conclusion is more than the result of a mere association of ideas which just happen to be about the same city. Talking about people as drawing inferences, the argument goes, is in part a matter of casting what they do in a causal explanatory framework. If I am to be said to be able to draw the required inference, there must a common causal factor in my thinking about Paris as the capital of France and my thinking that the Louvre is in Paris. Talk about this common causal factor is part of an explanation as to what makes it transparent to me that the city in question is one and the same, what it means to say that I am thinking about that city in the same way in both premisses, and can therefore arrive at that conclusion.

What I want to suggest, in short, is that talking about someone imitating somebody else is similarly in part a matter of casting what they do in a causal explanatory framework. Here, the question may be put as follows: in what way must what I see others doing and what I can do myself, respectively, be represented if I am to be capable of imitating what others are doing? And part of the answer is that there must be a common causal element to a subject's perception of the action performed by the other and her own production of that action, which explains why it is *that* action which is performed. Again, appeal to such a common causal factor is necessary to explain what makes it transparent to the subject how an observed action is to be replicated, how the subject can just read off from other's behaviour what to do. This, I take it, is the real force behind saying that perceived and performed actions are represented in the same mental code. However, it is also precisely the kind of claim to which those neurophysiological studies may be thought to lend support in which perception of others' actions selectively activates those brain areas also activated when the very same actions are carried out by the subject herself (see also Rizzolatti et al. 1996).

4. It is tempting to move from the idea of a common mental code, i.e. from the idea that perceived actions and performed actions are *represented* in the same way, to the idea that imitation involves *thinking* about others' actions in the same way as thinking about one's own actions. In my view, however, there are various grounds for finding such a move suspicious, or at least we have to be careful about how the notion of 'thinking about actions' is to be spelled out in this context.

Consider the case of deferred imitation. In experiments carried out by Meltzoff (1990), children saw an adult touching the top of a box with his forehead, whereupon a light came on inside the box. They saw this action only once, and the box was removed as soon as the adult had performed the ac-

tion, without giving the child the chance to touch the box herself. What Meltzoff found was that 14 month olds, when re-introduced to the box after one week's delay, reproduced the action they had seen the adult perform. Information about what the adult had done, it seems, was retained in such a way that it translated directly into doing the same thing.

One of the reasons why Meltzoff is interested in the outcome of this experiment is because he thinks it has important implications for theories of children's memory capacities. In this, I think, he is correct. What I think is not correct is to assume, as he does, that these studies provide evidence for a developing capacity for episodic memory, i.e. that the children can be said to remember the adult's action as a particular event in the past. What I wish to suggest, instead, is that the children remember something *about the box*, something they have been let in on by the adult, but something that they can remember without, say, having a concept of actions as individuated by the time at which they happen or by the person who carries them out. Imitation, here, is in the service of dealing with the world rather than dealing with others or oneself. In other words, the fact that it is someone else's act which is imitated need not enter into the intentions of the subject who is doing the imitating.

By saying this I do not wish to imply that there aren't serious questions to be asked about the kind of psychological understanding involved in deferred imitation. For instance, it may be argued that the capacity for deferred imitation develops together with a grasp of certain enabling conditions for actions, such as the fact that one must attend to the object and recognize that it affords a certain action. People, it might be said, must thus be understood as being different from other things in virtue of the fact that they possess particular causal powers needed to actualize certain affordances in objects. Similarly, it might be said that this understanding of people in general carries in it the seeds for a distinction between me and others. In this context, it might, for instance, be interesting to look more closely at the developmental stage at which the fact that somebody else does what the child is already capable of doing for herself becomes a source of frustration for children.

The main point I wish to stress here is that deferred imitation, inasmuch as it involves an understanding of others as psychological subjects, may not as yet involve a full understanding of their subjectivity, for instance in the sense that they may have desires that I don't share. For one thing, what drives the imitation, here, is the fact that what has been perceived had an outcome which, for the individual who imitates it, was desirable. It is difficult to see how an understanding of others' actions as potentially driven by desires that the individual herself does not share should develop from this basis.

5. Some theorists have argued that an important requirement for a full understanding of others' actions is the ability to *imagine* performing those actions (see, e.g., Barresi & Moore 1996). I think we can now see why imagination should be deemed important in this context. In imagination I can go through an action in my mind without the goal of that action being a goal that I desire. Perhaps it is by way of imagining doing what we see others doing that we come to have an understanding of them as agents like us, and yet agents whose subjective point of view is potentially different from ours.

One of the most interesting things shown by some of Decety's studies is that there is a very robust sense in which people can be said to go through an action in their minds when they imagine performing that action. Specifically, what we find is that, when someone imagines an action, the same brain areas are activated as when they actually carry out that action (Decety & Ingvar 1990; Decety 1996). We even find that changes in heart rate, respiratory rate, ventilation and blood pressure occur, which correspond to changes we would see if the action were actually carried out (Decety et al. 1993).

One thought that is sometimes put forward in this context is that imagination might provide me with an idea of what it would be like for me to perform the imagined action, such that, if I see someone else performing that action, I can somehow extrapolate from my own case to theirs. Imagination provides me, so to speak, with the subjective view that I would have if I were the agent, which I then impute to the other. This line of thought has obvious affinities with the approach to the problem of other minds suggested by the 'argument from analogy', and I think it faces the same sorts of problems as that approach. Very roughly speaking, if imagination revealed only what is true of my own case, it would be hard to see what could justify the thought that what is thus revealed to me might equally apply to someone else, or how I might even make sense of such an idea.

However, I believe that there is an alternative way of spelling out a role for imagination in the understanding of actions which avoids the pitfalls associated with the idea of an extrapolation from one's own case. Put bluntly, the suggestion would be that when I imagine performing the other's action, the question of what she is up to is somewhat beside the point. I am not so much after an understanding of what is going on with her, or what would be going on with me if I did the same thing as her. What I am after is an understanding of what it is to do it (I hope to make clearer what I mean by this in a moment).

There is one set of studies that Decety reports on which I think is particularly interesting in this context (cf. Decety et al. 1997; Grèzes et al. 1998). In these studies, positron emission tomography was carried out on participants

under two different subjective conditions. They were instructed to watch pantomimes, and in some cases they were told that they would be asked to recognize those pantomimes later, while in other cases they were told they would be asked to imitate them later. If I understood the data about the brain regions activated under these two conditions correctly, and the way in which they fit together with Decety's other studies, the results are consistent with the idea that participants in the imitation-condition actually imagined performing the action while they saw it, whereas participants in the recognition-condition typically didn't.

If this is correct, we have to ask just how imagination may be thought to come in here. Imagination is subject to certain causal constraints.[8] Consider a case in which I imagine performing an action that I have observed someone else performing. I think it is plausible to say that I can only be said to perform that action if my imaginative exercise is suitably causally related to what I have seen. There might not be such a causal connection, but then the best I could be said to be doing is imagining performing an action of a certain sort, where that sort of action just happens to be the same as the sort of action I have seen the other perform.

But there is also a further set of causal constraints that my imaginative exercise is subject to. This comes out when we consider the following kind of case. Suppose the person I am watching is a contortionist, bending her limbs in a manner far beyond my own physical capabilities. In this case, I might be able to bring myself to visualize what I might look like if my body was being contorted in the same way, or even imagine certain sensations I would have, such as a tension in my leg, or the feeling in my toes as they touch my hair. But, arguably, this falls short of imagining performing the action that the contortionist performs. In general, it seems that I can only be said to imagine performing a particular action if my imaginative exercise is suitably causally related to my own ability actually to perform that action. I might be able to push the limit somewhat, and imagine doing something that is not quite within the range of my current abilities, but beyond a certain point imagination simply gives out.[9]

If these considerations are along the right lines, how do they bear on the question as to what the participants in Decety's imitation-condition are trying to work out by using their imagination? As I have already indicated, my suggestion is that what they are trying to work out is what it is to perform the action in question. Now, if what I have been saying is correct, their imagination can only make available such knowledge in as far as they possess the abilities required for actually performing that action themselves, because the success of the imaginative project is dependent on those abilities being in place. But that

is just as it should be. After all, their knowledge ought to answer to what it is to perform the action, and it is difficult to see how imagination unfettered by ability could secure this.

What I have in mind by talking of knowledge as to what it is to perform a certain action is thus a species of knowledge 'from the inside', at least in the sense that it can't be acquired purely by observation, but presupposes certain physical capabilities on the part of the subject herself. But we must be careful not to conflate talking about knowledge 'from the inside', in this sense, with talking about knowledge pertaining only to me. While it is true that imagining performing an action presents that action from a subjective point of view, just like imagining seeing something presents what I imagine to see from a subjective point of view, it does not follow that this subjective view is presented to me as mine. What is presented is what it is (or would be) to perform a certain action or to see a certain scene, but not necessarily in such a way that I first understand these things for my own case and then extrapolate to that of others.

6. What I have done so far is talk about various ways in which we can be said to be responsive to others' actions, learn from them how to do things, or imagine actions, which do not seem to require the ability to individuate actions in terms of the agent by whom they are carried out. I wish to conclude with a speculation on the kind of context I believe we need to look at if we want to find out what it takes to have a concept of agency that involves conceiving of me as the author of some actions, and of others as the authors of other actions.

Take, for instance, the case of language acquisition. Without doubt, the ability to imitate others' utterances plays an important role in learning how to use language. But saying this tells us little about just how deep the connection goes. One idea here might simply be that speaking a language is a skill one can only learn by practising, like riding a bicycle. But, possibly in contrast to riding a bicycle, it is not a skill we could work out for ourselves. The only way we can acquire the relevant skill is by picking it up from others. Hence the importance of imitation. The basic thought here would be that language acquisition is a matter of acquiring a skill that involves following a convention, somewhat akin to learning to follow the highway code. Basically, we must grasp that communication can only succeed in so far as we and others abide by the same rules. It is possible to argue, however, that this view of language doesn't go quite deep enough.

What is left underdetermined by pointing out the conventional nature of language, it may be said, is the extent to which mature linguistic communication requires a grasp of others as language users like myself. It is one thing to

understand that communication is a means of doing certain things which relies on others' co-operation. It is another thing to understand that communication is a means of doing certain things that can only succeed in so far as it is also at the other's disposal.

That there is a difference here might come out more clearly when we consider the difference between imperatives and declaratives. The latter difference seems to turn, in part on the grasp of others as language users that is involved in using imperatives and declaratives, respectively. Arguably, it is from others that we must first learn the use of both imperatives and declaratives. At its most primitive, however, the actual use of an imperative does not seem to require grasp of the other as a potential issuer of imperatives himself. When I am using an imperative, I might be interested in what the other can or can't do only to the extent that I am interested in what I can make them do. In some ways, I treat the other like an extension of myself. In using a declarative, by contrast, what the other can or can't do seems to enter into my reasoning in a stronger sense. Put crudely, what I must grasp is that I need to choose the same words that the other would use to refer to the things that I am trying to tell them about. In other words, when using declaratives, I must be sensitive to the fact that the success of my project is not just dependent on what I can *make* the other do, but also on what they can do *independently of me*.

The more general suggestion I wish to put forward is that a mature grasp of oneself and others as agents is interdependent with grasp of the communicative use of action and imitation. Admittedly, this is as yet extremely vague, and much more would need to be said about how this suggestion is to be spelled out in detail. The basic point I wish to make here is that, in as far as communication in this sense requires the ability to imitate, to do what we have seen others doing, what we are considering here is not just a new kind of context, different from the ones discussed previously, in which imitation is used. We are also considering a new kind of imitation. If what I have been saying is right, imitation, in the sense required here, is not just a matter of doing what we have seen others doing (where the fact that we have seen them doing it causally explains our capacity to do it), it is also a manifestation of our grasp that what we are doing is something they are capable of doing themselves.

Notes

1. The distinction that seems to be at work here shares much with the distinction, introduced by Daniel Dennett (1969), between a sub-personal and a personal level of explanation.

2. See also Prinz (1997).

3. To be sure, neuroscience might also have a say on the latter question, by investigating what it is about our brains that allows us to conceive of our own actions as ours and those of others as theirs. Decety, for instance, discusses different mechanisms for discriminating self-produced from observed action as proposed by Frith (1992) and Jeannerod (1999).

4. Influential accounts of a general connection between perception and action can be found in Sperry (1952) and Gibson (1979). The term 'common mental coding' is also often used in this context. Thus, for instance, it has been argued that perception and action use the same system of egocentric spatial co-ordinates (Evans 1982). Arguably, however, the idea of a common mental code for observed and performed actions, as outlined above, is more than just a specific application of this general idea. In what follows, I shall therefore use the term in the narrow sense intended, for instance, by Meltzoff and Moore.

5. There is some evidence that suggests that the ability to recognize biological motion is, at least in one sense, part of a more general ability to recognize structure from motion, which, for instance, also allows us to infer the shape of inanimate objects in rotation from the relative movements their parts. Thus, for instance, in a patient studied by Vaina et al. (1990), performance on early motion tasks was very poor, but both biological motion and structure from motion perception remained intact. See also Marcar et al. (1997).

6. This example was brought to my attention by Johannes Roessler.

7. One way in which this point can be brought out is that talk of an individual as imitating another individual also has specific counterfactual implications, viz. that the individual doing the imitating would have done the same as the individual who was being imitated, even if the latter had done something different.

8. In what follows, I draw heavily on Williams (1973) and Peacocke (1998).

9. Decety and Jeannerod (1996) report interesting evidence which suggests that subjects need not conceptualize the constraints their imaginative project is answerable to in order to be sensitive to such constraints in going through the imaginative exercise. They demonstrated that participants who were instructed to imagine walking through a gate presented at a certain distance from them performed differently depending on the width of the gate. Their responses were found to accord with Fitt's law, according to which there is a trade-off between speed and accuracy in action. Yet, it is unlikely that participants knew about this law.

References

Barresi, J. & Moore, C. (1996). Intentional relations and social understanding. *Behavioural and Brain Sciences, 19*, 107–154.

Davies, M. (1991). Concepts, connectionism and the language of thought. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Decety, J. & Jeannerod, M. (1996). Mentally simulated movements in virtual reality: Does Fitts's law hold in motor imagery? *Behavioural Brain Research, 72*, 127–134.

Decety, J. & Ingvar, D. H. (1990). Brain structures participating in mental simulation of motor behavior: A neuropsychological interpretation. *Acta Psychologica, 73*, 13–34.

Decety, J. (1996). Do imagined and executed actions share the same neural substrate? *Cognitive Brain Research, 3*, 87–93.

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*, 1763–1777.

Decety, J., Jeannerod, M., Durozard, D., & Baverel, G. (1993). Central activation of autonomic effectors during mental simulation of motor actions in man. *Journal of Physiology, 461*, 549–563.

Dennett, D. C. (1969). *Content and consciousness*. London: Routledge & Kegan Paul.

Evans, G. (1982). *The varieties of reference.* Oxford: Oxford University Press.

Fadiga, L., Fogasi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Frijda, N. (1969). Recognition of emotion. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 4.* New York: Academic Press.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia.* LEA.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 593–609.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Grèzes, J., Costes, N., & Decety, J. (1998). Top-down effect of the strategy on the perception of biological motion: a PET investigation. *Cognitive Neuropsychology, 15*, 553–582.

Jeannerod, M. (1999). To act or not to act: perspectives on the representation of actions. *The Quaterly Journal of Experimental Psychology, 52A*, 1–29.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14*, 201–211.

Marcar, V. L., Zihl, J., & Cowey, A. (1997). Comparing the visual deficits of a motion blind patient with the visual deficits of monkeys with area MT removed. *Neuropsychologia, 35*, 1459–1465.

Meltzoff, A. N. (1990). Towards a developmental cognitive science: The implications of cross-modal matching and imitation for the development of representation and memory in infancy. In A. Diamond (Ed.), *Annals of the New York Academy of Sciences: Vol. 608. The development and neural bases of higher cognitive functions.* New York: New York Academy of Sciences.

Meltzoff, A. N. & Moore, M. K. (1995). Infants' understanding of people and things: from body imitation to folk psychology. In J. L. Bermúdez, A. Marcel, & N. Eilan (Eds.), *The Body and the Self.* Cambridge, MA: MIT Press.

Peacocke, C. (1998). Conscious attitudes, attention and self-knowledge. In C. Wright, B. C. Smith, & Cynthia McDonald (Eds.), *Knowing our own minds*. Oxford: Oxford University Press.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129–154.

Rizzolatti, G., Fadiga, L., Galese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141.

Roth, M., Decety, J., Raybaudi, M., Massarelli, R., Delon, C., Segebarth, C., Gemignani, A., Decorps, M., & Jeannerod, M. (1996). Possible involvement of primary motor cortex in mentally simulated movement: a functional magnetic resonance imaging study. *NeuroReport*, *7,* 1280–1284.

Sperry, R. W. (1952). Neurology and the mind-brain problem. *American Scientist, 40*, 291–312.

Williams, B. (1973). Imagination and the self. In *Problems of the Self*. Cambridge: Cambridge University Press.

Vaina, L. M., LeMay, M., Bienfang, D. C., Choi, A. Y., & Nakayama, K. (1990). Intact biological motion and structure from motion perception in a patient with impaired motion mechanisms: A case study. *Visual Neuroscience, 5*, 353–369.

Chapter 4

The scope and limits of mental simulation

Pierre Jacob
Institut Jean-Nicod, CNRS, Paris

1. Mental simulation: A great idea

The goal of the present paper is to examine the scope and limits of an approach to mindreading based on a heuristic called "mental simulation". In the present introductory section, I shall distinguish between two versions of the approach to mindreading based on mental simulation: an introspectionist version and an anti-introspectionist version. In the second section, I shall examine empirical evidence from cognitive neuroscience in favor of the view that mental simulation plays an important role in human cognition. Drawing on empirical evidence, I will argue that mental simulation has four fundamental features: firstly, it is a general purpose heuristic. Secondly, it is motivated by the rejection of the picture of mindreading as a capacity for detached theorizing. Thirdly, part of the idea of mental simulation can be captured by the notion of pretence. Fourthly, another part of it can be captured by the notion of mental mimicry. In the third section of the paper, I will examine the so-called "off-line" model of the prediction of another's decision. In the fourth section of my paper, I will examine problems faced by the introspectionist version of the mental simulation approach to first-person mindreading. In the fifth and last section of the paper, I will provide reasons for being sceptical about the view that mental simulation could provide a full account of third-person mindreading. I will first argue that there is a basic difference between engaging in pretence and using pretence for representing the thoughts of others. Finally, I will argue that representing the thoughts of others cannot be captured by mental mimicry.

The idea of mental simulation is a great idea with a great potential. It is currently under intense scrutiny in various areas of cognitive science: in developmental cognitive psychology as well as in cognitive neuroscience. In cognitive neuroscience, the idea has lately emerged that the ability to simulate an action

is crucially involved in the ability to perform it. In developmental cognitive psychology, it is suggested that perhaps our grasp of mental concepts depends similarly upon an ability to engage in pretence or imagination.[1]
indexImagination Finally, in various areas of cognitive science, mental simulation is explicitly linked to mimicry, the ability to engage in the mimicking (or duplication) of the behavior of others.

In developmental cognitive psychology, there is presently a lively discussion about the cognitive structure and nature of our ability to understand mental concepts and to ascribe propositional attitudes to ourselves and to our conspecifics. Roughly speaking, there is a major division between so-called *theory*-theorists and *simulation*-theorists of what I will call naive psychology. Theory-theorists can in turn be divided into two groups: some, like Wellman (1990), Gopnik and Meltzoff (1997), Gopnik and Wellman (1992, 1994) advocate a *child-scientist* version of their view. Others, like Leslie (1994a, 1994b, 1995) and Baron-Cohen (1994, 1995), advocate a *nativist domain-specific* version of the Theory Theory. Theory-theorists of both persuasions depict the human mastery of mental concepts (i.e., the ability to explain intentional actions in terms of beliefs and desires) as some kind of tacit knowledge. They derive inspiration from the Chomskian study of the language faculty where the ability of the speaker of a natural language to produce and understand an infinite set of grammatical strings of words of his or her language is explained in terms of his or her knowledge of the grammar of the language. Simulation theorists too fall roughly into two groups: some, like Goldman (1989, 1993), Heal (1986, 1995) and Harris (1992), the *introspectionist* simulation theorists, give priority to *first-personal awareness* of mental states in an individual's psychological knowledge. Others, like Gordon (1986, 1996), an *anti*-introspectionist simulation theorist, deny such priority. In fact, simulation theorists of both persuasions reject the picture of the human ability to interpret other human beings as reflecting some underlying *detached* knowledge. According to them, interpretation of others involves primarily an ability to *do* something, i.e., the ability to replicate the other within oneself.[2]

Presumably, unlike "core" knowledge of language, as Chomsky (1975) understands it, or unlike naive physics, naive arithmetic and naive geometry, as developmental psychologists such as Carey and Spelke (1994) understand it, mental simulation is not a domain specific informationally encapsulated cognitive capacity. Like memory and like the imagination, mental simulation is presumably a general purpose cognitive ability or heuristic. For example, it is probably not restricted to entertaining possible non-actual states of affairs per-

taining to human action and human mental states. It probably can be involved in considering physical, astronomical or chemical counterfactual situations.

In the context of an examination of the fundamental conceptual structure of naive physical thinking, Campbell (1994) distinguished two broad ways of thinking of the causal significance of various spatial and temporal notions: a practical grasp or an engaged way of thinking and a theoretical or detached understanding. In the present context, this contrast can help us capture the basic disagreement between so-called theory-theorists and simulation theorists of naive psychology. At least, it helps one understand the perspective of the simulation theorists on the debate: simulation theorists think that a human being's grap of mental concepts relies, not on a capacity for detached theorizing, but on a capacity for engaging in the impersonation or the replication of the mental life of one's conspecifics. This is, I take it, what Gordon (1996: 11) has in mind when he contrasts what he calls a "cold methodology" (i.e., "a methodology that chiefly engages our intellectual processes") and what he calls a "hot methodology" (one "which exploits one's own motivational and emotional resources").

2. Mental simulation, motor imagery and the preparation of action

I will start with an investigation of the process of mental simulation in the area where, I think, it belongs primarily, namely the ability to perform motor actions. Research in cognitive neuroscience has demonstrated that visual imagery shares some neural resources and mechanisms with visual perception. If true, then the processes of visual imagery would be parasitic on visual processes and mechanisms. On this basis, it has been suggested e.g., by Currie (1995) that the exercise of visual mental imagery consists in the mental simulation of vision. On this view, in visual mental imagery, the visual system would be, in Stich and Nichols' (1992, 1995) useful phrase, "driven off-line": it would be disconnected from its normal or standard inputs and outputs. Similarly, in the cognitive neuroscientific study of the motor system, several lines of investigation support the idea that motor imagery and the execution of motion share some neural resources and mechanisms. On this basis, it has been suggested (e.g., by Jeannerod 1994, 1997) that the ability to simulate a movement is crucial to the capacity to perform it. Indeed, the only difference between performing and simulating a movement would be that, in the latter case, execution is inhibited or cancelled.[3] I will summarize some of the evidence for the view that imag-

ining a movement – 'imaging' or simulating it – and performing it have much in common.

Intuitively, it does seem as if mental simulation of motion is at work in the learning of motor skills both on the part of the student and on the part of the teacher. During a violin or a tennis lesson, at some point, the teacher may execute a particular motor sequence. While watching the teacher, the student will presumably mentally rehearse the movement before executing it. Conversely, while examining the student, the teacher will replicate within himself the appropriate motor sequence without performing it then. Sport addicts watching sport (e.g., tennis) on TV are said to mentally execute the action they are observing without physically performing it. Indeed, they seem to directly experience frustration when the athlete whom they are watching misses his target.

Psychophysical measurement has revealed that an actually or overtly performed action involves the same duration as its mental or implicit counterpart. This is true in the domain of learned skills where comparison has been made between overt and mental recitation of the alphabet or in comparing an actual writing task to its mental counterpart. And it is true of natural motions: walking to a specified target requires the same time as imagining walking to the same target.[4]

Sport psychologists have reported that mental practice facilitates future performance. Increase in muscular strength consequent upon actual training (i.e., undergoing actual muscle contraction) has been compared with increase in muscular strength consequent upon mental training. In mental training, the subject imagines muscle contraction while the muscle is quiescent. It has been found that muscular strength increased by 30% upon actual training while it increased by 22% upon mental training. In both actually performed and simulated physical exercises, heart rate and respiration rates were measured and found to increase.[5]

What Jeannerod (1994, 1997) calls "motor representations" are involved in the preparation of action. When motor preparation is smoothly followed by execution, the contents of motor representations are not available to the agent's consciousness. The contents of motor representations do become available to consciousness when, for one reason or another, the execution of action is cancelled. The contents of motor representations then surface to consciousness as motor imagery. Hence, motor imagery is motor representation under inhibition of execution.

The content of a motor representation combines two broad ingredients: a representation of the goal of the action and a representation of the body in

action. The latter encodes information of roughly three kinds: (i) about the expected duration of the action; (ii) about the amount of force and effort to be invested in the execution of the action; and (iii) about the "motor rules", i.e., the kinematic regularities and biomechanic constraints on bodily motions. When the action is visually directed towards an object (as in grasping), then information about the visual attributes of the object is processed via what Jeannerod calls a "pragmatic" mode of visual information-processing, as opposed to a "semantic" mode of visual information-processing. There is a lively discussion as to how this functional opposition between modes of visual information-processing can be mapped unto Ungerleider and Mishkin's (1982) anatomical distinction between the dorsal stream and the ventral stream in the visual system.[6] Roughly, perceptual representations of the visual attributes of objects delivered by the semantic processing of visual information provide food for conceptual thought. Motor representations of visually perceived objects delivered by the pragmatic system provide visual information about the object as a goal of action, much as an affordance, to use Gibson's celebrated term. It has sometimes been suggested that the semantic system of information-processing provides perceptual representations of an object in an object-centered or allocentric frame of reference, whereas the pragmatic system could only provide motor representations of the same object in an egocentric frame of reference. This must be a mistake, though, because a grasping task involves both the transportation of the hand at the appropriate location of the object in space and the appropriate scaling of finger grip required to seize the object. The former does require spatial information about the object coded in an egocentric frame of reference. But the latter requires visual information about the size, shape and relations between parts of the object. And this must be coded in an object-centered frame of reference.

Finally, I will mention one last piece of neuroscientific evidence in favor of the idea that the ability to simulate a motion is crucial to the ability to perform it: this is the celebrated discovery of the so-called "mirror neurons" in the premotor cortex of macaque monkeys by Rizzolatti et al. (1995), which Gallese and Goldman (1998) have made much of in a recent paper on mental simulation. Rizzolatti et al. (1995) found in premotor area F5 a class of neurons with two remarkable properties. First, they fire when the monkey is performing a certain kind of goal-directed visually guided hand motion. They are sensitive not to purely cinematic or dynamic features of hand motion but to the action itself, where the action involves both the specification of a motor act and a goal. These neurons are highly selective to the kind of finger configuration involved in motion. So Rizzolatti et al. (1995) classify such neurons as "grasping

neurons", "holding neurons", "tearing neurons", "manipulation neurons". Some grasping neurons discharge during finger extension, others during finger flexion. Mirror neurons show selectivity to different kinds of hand prehensions: some holding neurons discharge when a sphere is being held, others when a cylinder is being held. Secondly, not only do these mirror neurons fire when the monkey actually performs one of these visually guided hand motions, they also fire when the monkey is not doing anything and is merely watching either another monkey or an experimenter perform the very same action. However, when the monkey is not doing anything and is merely watching, mirror neurons will fire only if the experimenter grasps a piece of food. If the experimenter makes a grasping movement without food or if food is grasped with a tool, mirror neurons will not fire.

Several writers have linked the notion of mental simulation to the notion of *mimicry*. As Gordon (1986, 1996) and Goldman and Gallese (1998) emphasize, when I engage in pretence for the purpose of reading the mind of someone else, I engage in a process of mimicking the other by creating pretend states within myself. On this view, reading another's mind consists at bottom in the creation within oneself of mental states that are copies, reproductions, replications or duplications of the other person's mental states. Mirror neurons may be one among several interesting mechanisms for mimicry recently discovered in humans by cognitive science. Now, mimicry consists in the duplication of states across different organisms. Emotional contagion is an instance of mimicry. So are laughing and yawning. Meltzoff and Moore's (1983) observations show that a few minutes after birth, a human infant can imitate facial gestures of mouth opening and tongue protrusion. In the words of Gallese and Goldman (1998), "mirror neurons appear to form a cortical system that matches observation and execution of actions". Clinical evidence has revealed that so-called "imitation patients" with prefrontal lesions (first observed, I believe, by Lhermitte) compulsively reproduce goal-directed actions performed in front of them. Echopractic patients compulsively imitate gestures performed in front of them by the experimenter. As reported by Gallese (2000), in these patients, the normal inhibitory mechanisms which block the execution of motor plans triggered by observation have been impaired by the prefrontal lesion.

3. Mental simulation: The off-line model for predicting another's decision

The thought of extending the concept of mental simulation from the motor system to higher cognitive functions (such as naive psychology) has been entertained by Gordon (1986: 70), who has explicitly linked the ability to engage in the replication of the mental life of others to the ability to anticipate, without executing, others' motor behavior. Now, one of the earliest appearances of the idea of mental simulation outside the study of the motor system is Kahneman and Tversky's (1979) appeal to the "simulation heuristic" as an explanation of the ability to predict other human beings' emotions. This study has been much discussed by Goldman (1989). Kahneman and Tversky (1979) provided subjects with a scenario in which two persons A and B are driving to the same airport to catch two different planes scheduled to depart at the same time. Both miss their flights. But unlike A's plane, B's plane has been delayed so that B almost but not quite made it to his flight. The question asked is: Who is more upset? A or B? 96% of the subjects responded that B is more upset than A. On Kahneman and Tversky's (1979) view, subjects provide a prediction by imagining or simulating the travellers' predicament: they engage in pretence, they put themselves in A's and B's shoes; they compare A's and B's experiences; and as a result, they predict that B is more upset than A.

Now, being upset is an emotion. Although it is perhaps controversial, I assume that emotions differ from what I would like to call "visceral states" (such as hunger, thirst, the urge to sleep or to urinate) in that emotions, unlike purely visceral states, have intentional objects and they must be triggered by beliefs. Unlike beliefs though, emotions in turn have typical physiological effects such as nausea, heart-rate acceleration and deceleration, lumps in the throat and so on.[7]

One of the clearest models of the process of mental simulation outside the motor system has been provided, I think, by Stich and Nichols (1992, 1995): it is the off-line simulation model of predicting somebody else's decision. On this view, a person uses his or her own decision-making (or practical reasoning) system. When running on line, my decision-making system takes as inputs my own beliefs and desires and generates as output my decisions to act. This very same system may be taken off-line and used for the prediction of somebody else's decision. In which case, my decision-making system is disconnected from its standard inputs and outputs. When taken off-line, it takes as inputs pretend-beliefs and pretend-desires – beliefs and desires which are presumably held by the person whose decisions I am predicting. And it outputs a pretend-decision.

When I use my decision-making system for the purpose of taking a decision, I decide to act on the basis of my beliefs and desires, and finally I act. When I take my own decision-making system off-line to predict what somebody else will decide to do and when I give it pretend-beliefs and pretend-desires, I do not end up with a decision to act on the basis of these pretend-beliefs and pretend-desires. A fortiori, I don't end up acting. Rather, I store the conclusion of my decision-making system taken off-line (my prediction or my pretend-decision) into my "belief-box".

Decisions to act are, of course, fundamental steps in the preparation of intentional action. And they are, I think, closely linked to emotions. It is, I think, worth pondering over the fact that the psychology of reasoning suggests that there are two broad ways in which the rationality of a reasoning process can be disturbed in normal subjects. On the one hand, theoretical reasoning (either of a deductive or of an inductive sort) leading to beliefs (or theoretical conclusions) gives rise to various fallacies which are prompted by purely cognitive illusions. In deductive reasoning, the fallacy of the negation of the antecedent or the fallacy of the affirmation of the consequent are of this kind. So are the biases in judgments under uncertainty described by Kahneman and Tversky in the domain of inductive reasoning. On the other hand, in the domain of practical reasoning, there are well-known phenomena such as wishful thinking or akrasia where emotions may disturb an evaluation procedure or a decision process. In the former case, I may deceive myself about a person I am in love with by underestimating evidence to the effect that my feelings for her are not reciprocated. In the latter, I may believe that everything considered it is better for me not to smoke; however, I cannot refrain from smoking. In normal subjects, emotions may thus interfere with the rationality of practical reasoning, i.e., reasoning leading to intentional action via a decision procedure.

Surprisingly, as work reported by Damasio (1995) suggests, *lack* as opposed to overflow of emotional responses seems to interfere dramatically with the ability to take personal decisions. Patients with frontal lobe lesions (such as Phineas Gage) become pathologically detached: they lose the ability to experience the emotions which normal subjects experience in the decision-taking process. As a result of this disengagement from normal emotional response, they seem to lose the ability to take rational decisions and to make beneficial choices. Now, taking a decision to act and making a choice involve the ability to commit oneself to one branch of an alternative, stick to it and give up the other branch of the alternative. This is what, it seems, frontal lobe patients examined by Damasio miss. Interestingly, Damasio compares such frontal lobe patients to anosognosic patients who, after a stroke, are paralyzed on the left side of

their body and unaware of their condition.[8] Not only are anosognosic patients paralyzed, but they have lost proprioceptive access to the paralyzed side of their body. Frontal lobe patients cannot take practical decisions because they are deprived of the ability to experience the normal emotions prompted by personal decision problems.

As Gordon's contrast between "hot" and "cold" metholodogies indicates, the role of emotional responses in the process of decision-making and the preparation of intentional action has been emphasized by advocates of the simulation theory of naive psychology when they characterize mental simulation as a process of *empathy*: the main target of an empathetic process is the replication of somebody else's emotions. Now, in using the simulation heuristic for predicting the travellers' emotions and in using one's own decision making-system off-line to predict another's decision as well, one takes pretend-beliefs for granted. In other words, in both tasks, one runs the simulation process on the basis of pretend-beliefs: the simulation process itself does not yield the beliefs in question.

So the question I want to raise in the sequel is: To what extent can the mastery of a fundamental mental concept such as the concept of belief be accounted for by mental simulation? Can mental simulation yield a constitutive account of beliefs or of the possession conditions for the concept of belief? We typically deploy or use our concept of belief in belief-ascriptions, i.e., when we ascribe a belief to another person. Now, there are two basic ways a simulation-theorist might want to provide a positive response to the question of whether the process of mental simulation can explain our full mastery of the concept of belief or yield a constitutive account of the possession conditions for the concept of belief. One is the introspectionist view argued for by Goldman (1989, 1993) in several papers. The other is Gordon's (1986, 1996) anti-introspectionist version of the simulation theory. I shall start with the former.

4. The introspectionist version of mental simulation and first-person mindreading

On Goldman's (1989, 1993) introspectionist version of the simulation theory, we are endowed with a first-personal mastery of concepts of propositional attitudes. We recognize our own beliefs and desires from a first-person perspective because they have intrinsic qualitative properties attached to them just like sensations and emotions do. In other words, the introspectionist version of the simulationist account presupposes that there is a phenomenology associ-

ated to believing. Then, using ourselves as a model for others, we can ascribe propositional attitudes to others.

So the introspectionist version of mental simulation has two major ingredients: first of all, it assumes that we derive knowledge of others' psychological properties from a prior knowledge of our own psychological properties. Secondly, it provides an account of first-personal knowledge of our own psychological properties. In the present section, I shall accept the first thesis that we use ourselves as a model for deriving knowledge of others' psychological properties.[9] And I shall examine two possible ways an advocate of the introspectionist version of simulation theory could account for first-personal knowledge of our own mind. The question is: Can mental simulation provide an account of psychological self-knowledge or first-person mindreading? On my view, Goldman (2000) has provided a decisive reason why it cannot. To see why not, let us go back to the off-line model of my prediction of somebody else's decision. Suppose I am trying to predict a decision you are about to make and I am using the pretence-based simulation routine. On this model, I create in myself some pretend beliefs and desires that are supposed to match your beliefs and desires – those beliefs and desires on which you are basing your decision (which I am trying to predict). I feed these pretend states into my practical reasoning system; I turn it on; and it supplies a pretend-decision. I then use this pretend-decision to predict your decision. So, on this model, using a simulation pretence-theoretic strategy to do third-person mindreading depends on or presupposes the ability to do first-person mindreading.

But now the question naturally arises: How did I detect the output of my own decision-making system? How did I determine which decision was selected by my practical reasoning system on the basis of the pretend beliefs and desires which served as inputs? Why could I not use the pretence-theoretic simulation heuristic to respond to this question and treat myself in a third-personal way? The reason why I cannot is the following. Were I to use simulation to identify the output of my own decision-making system, I would be involved in an infinite regress and I would never succeed in the task. In order to identify the result of my practical reasoning system, I would have to use a second-order simulation heuristic. This second-order simulation heuristic would generate a second-order output. Now, if I need a second-order simulation heuristic to determine the output of my first-order simulation, then presumably, I could not identify the second-order output of the second-order simulation unless I used a third-order simulation heuristic, which would generate a third-order output, and so on endlessly. But presumably, I can iden-

tify the output of my first-order simulation heuristic. And so, I do not use the simulation heuristic to read my own mind.

So the question is: How do I read my own mind? What underlies direct immediate introspective self-knowledge? Goldman (2000) has sketched an answer to this question. According to him, we have what he calls "an inner detection mechanism" attuned to identifying our own mental states. In the past, Goldman (1989, 1993) has argued that propositional attitudes (beliefs and desires), as well as sensory states, have so-called *phenomenal* properties, i.e., intrinsic qualitative properties. If they do, then these phenomenal properties would presumably be the targets of the inner detection mechanism.

First of all, I want to point out that Goldman's proposal will not get off the ground unless beliefs and desires *do* indeed have phenomenal properties. I am sceptical. For example, I doubt very much that there is something it is like to believe that 5 is a prime number. I think that beliefs can give rise to emotions and emotions do have intrinsic phenomenological properties. But the idea that beliefs have intrinsic phenomenological properties might arise from a confusion between a belief and the belief-caused emotions. Plainly, one and the same proposition held under the same attitude (i.e., belief) can give rise to different emotions in two different persons at the same time in virtue of the differences between their belief systems and their values. Think for example of the proposition that Bill Clinton was not impeached in the Fall 1998. Presumably, both Kenneth Starr and Bill Clinton can hold the belief that this proposition is true. However, if they do, this single belief will not prompt the same emotional response in both.

I will now briedfly sketch my general reasons for scepticism about Goldman's (2000) inner detection mechanism of our own psychological properties. In general, I assume that awareness of anything comes in two varieties: conceptual and non-conceptual. Conceptual awareness is propositional or factual awareness. In the simplest case, it is awareness of the fact – i.e., belief – that some particular object exemplifies a given property (or that an n-tuple of objects exemplifies an n-ary relation). Non-conceptual awareness, by contrast, is simply awareness of a particular object. Conceptual awareness is belief. Non-conceptual awareness arises from simple (non-epistemic) perception. It may look as if one could not be aware of an object without being aware of any of its property. This may be true, particularly if you have a visual experience of an object. However it is also true, I believe, that you may touch an object (without seeing it), be aware of it and have simply no idea what it is. Furthermore, there is, I think, no particular property of an object of which one must be aware in order to be aware of the object. It may be misleading to say that a person e.g.,

saw something if the person utterly failed to identify it. Although it may be misleading to say so, it may be perfectly true. You can therefore be aware of an object without being aware of some of its properties. An akinetopsic patient, for example, may be visually aware of a moving object in virtue of the fact that he is aware of its color and shape without being aware of the motion of the object and without being aware that it is moving. A visual form agnosic patient may be visually aware of an object without being aware of its shape and/or size.

I further assume that a self – any self – exemplifies two kinds of properties: bodily properties and mental or psychological properties. Consider first representations of one's body. Representations of one's body (or bodily self-awareness) can be either non-conceptual or conceptual. Non-conceptual bodily self-awareness may involve proprioceptive representations of the position and motions of one's bodily parts (one's limbs, one's head, one's eyes). Proprioceptive representation is a central kind of perceptual representation. One can also entertain conceptual representations of one's body in the form of various theoretical anatomical and physiological beliefs about the structure, nature and function of one's body parts.

However, when it comes to mental self-awareness, I accept Shoemaker's (1996) famous criticism of the Perceptual Model of Introspection. On this assumption, there is no choice: one cannot have perceptual (or sensory) non-conceptual self-awareness of the content of one's own mind. One can only form beliefs about the content of one's experiences, beliefs and desires. One's sensory experiences of environmental objects and properties are paradigms of perceptual non-conceptual awareness. But one can no more directly perceive one's sensory (or perceptual) experiences than one can perceive one's own beliefs and desires. One can only entertain conceptual representations of what one experiences, one believes and one desires. Awareness of one's own mind is bound to take the form of higher-order beliefs about the self. One cannot know what one believes unless one forms a higher-order belief about one's lower-order beliefs. For example, I know that I believe that George W. Bush is the new president of the US because I can form the higher-order belief that I believe that George W. Bush is the new president of the US – much as I deplore it. Goldman's (2000) inner detection mechanism, I suggest, wrongly assumes that like bodily self-awareness, mental self-awareness comes in two varieties: conceptual and non-conceptual. He wrongly assumes, I think, that one can sense – or have sensory representation of – one's own beliefs.

5. The anti-introspectionist version of mental simulation and third-person mindreading

The anti-introspectionist simulation account of our mastery of concepts of propositional attitudes is in a way more radical than the introspectionist account. It is held by Gordon (1986, 1996) in various papers. The idea is that when I deploy the concept of belief in ascribing a belief to somebody else, I engage in pretence: I mimick the other person by going into a pretend-state. As Gordon (1986: 68) himself puts it: "to attribute a belief to another person is to make an assertion, to state something as a fact, *within the context of practical simulation*. Acquisition of the capacity to attribute beliefs is acquisition of the capacity to make assertions in such a context". I will first discuss some important difference between engaging in pretence and using pretence for the purpose of representing the beliefs of another person. Then I will provide some reasons for thinking that attributing beliefs to somebody else cannot be identified to a process of mimicry.

5.1 Pretending and representing beliefs

Gordon's (1986, 1996) account derives from Quine's (1960) famous remark to the effect that indirect quotation and the attribution of propositional attitudes are "essentially dramatic acts" in which we "cast ourselves in unreal roles" and "we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then we say what, in our language, is natural and relevant for us in the state thus feigned". According to Stich (1983: 84–88), who elaborated on Quine's insight, when a person ascribes a belief to somebody else, she does *two* things at once: first, she "performs a little skit, doing a bit of play acting" in which the 'that'-clause "constitutes the script for the play". Secondly, to say that subject *S* believes that *p* is to say that *S* is in a state similar to the state which would play the same typical causal role if my utterance of sentence "p" had had a typical causal history. In other words, to say that *S* believes that *p* is to say that *S* is in a state similar to the state I would be in were I to utter "p" in normal conditions. In other words, Stich does not present the first part of his account as a full account of belief-ascription or as a constitutive account of our mastery of the concept of belief.

Now, the crucial difference between Stich's and Gordon's account is that the former is a *two*-step account, whereas the latter is a *single*-step account. In other words, Gordon simply leaves out the second part of Stich's account. In the second part of his dualistic account, Stich appeals to a notion of similarity

between belief contents and belief states in his explanation of belief-ascriptions. He thus seems to concede that the ability to engage in pretence or "to perform a little skit" will not suffice to deliver the concept of belief. But, as noticed by Fuller (1995:26), on Gordon's account, in ascribing a belief to another person, I merely perform a simulation which culminates in a pretend assertion. As Fuller (*ibid.*) puts it, on Gordon's account, "we are left hanging": we don't know what the belief-ascriber is supposed to do with her pretend assertion. Gordon thus presupposes that once the pretend assertion has been made, the job of the belief-ascription is completed. On his view, there is nothing more to our mastery and deployment of the concept of belief than engaging in pretence, whether in pretend assertion or in pretend thought. And this, I think, is the fundamental weakness of Gordon's simulation account of our concept of belief, as I want to show succinctly now.

As I said earlier, all simulation theorists of naive psychology insist that our ability to interpret others is at bottom the ability to engage in pretence, i.e., the ability to replicate the mental life of others. However, I do not think that belief-ascriptions are *pretend assertions* or *pretend thoughts*. I think that belief-ascriptions (whether utterances or thoughts) are genuine thoughts and/or full assertions with truth-conditions: they are true or false. But they have a characteristic complexity: to think that someone else thinks that *p* is more complex than to think that *p*. And plausibly, unless some can think that p, she cannot entertain the thought that someone else thinks that *p*. It is one thing to engage imaginatively in pretence. It is something else to describe somebody else's state of mind. If and when I ascribe to a person a belief whose content I take to be false, I can pretend that the belief is true. I can thus imagine what the world would have to be like for the belief to be true. However, when I correctly ascribe a false belief to someone (as when I think or say e.g., that Maria believes that witches have magical powers – something I myself don't believe), I state a fact about the actual world. It may be true in Maria's belief-world that witches have magical powers. I can pretend that witches have magical powers for the purpose of exploring Maria's belief-world. If I think to myself or utter the sentence "witches have magical powers", this might be a pretend thought or a pretend assertion. But when I ascribe to Maria the belief that witches have magical powers, I entertain a complex thought one ingredient of which is the concept of belief. This complex thought which contains the concept of belief is not a pretend thought: it describes a portion of the actual world which contains Maria and her beliefs.

So, I literally disagree with Gordon's claim that to make a belief-ascription is to make an assertion (or to entertain a thought) *within the scope of a pretence*

or a practical simulation. Ascribing a belief is like describing a piece of fiction or stating a fact about it. One thing is to enjoy a piece of fiction by engaging in pretence and imagining that e.g., Sherlock Holmes was a smart pretentious detective who lived 221 Baker street in London. Another thing is to say or think that many of Conan Doyle's works depict Sherlock Holmes as a smart pretentious detective who lived 221 Baker street in London. In the former case, my pretend thought is about a possible world containing a fictitious character. My latter thought is about the actual world: it is about an actual piece of fiction (written by a real man whose name was Conan Doyle) which anyone can read and enjoy – by engaging in pretence. Conan Doyle's work belongs to the actual world and it depicts a fictional world to which Sherlock Holmes belongs. Belief states are like works of fiction. One may engage in simulation or pretence to characterize their content. But when one states facts about either works of fiction or about beliefs, one is *not* engaging in pretence. One is describing portions of the actual world.

One could perhaps, as Recanati (1997, 2000) has suggested, use Perry's (1986) distinction between what a representation *concerns* and what it is *about* to capture the semantic distinction between what is going on in pretence and what is going on when one exploits pretence to describe somebody else's beliefs. When, being in Paris, I think or utter the sentence "it's raining here" or "it's raining in Paris", my thought or utterance is about Paris and it may concern some larger situation (where a contrast between Paris and another city may be relevant). I'm entertaining a weather-thought with a constituent for the place: "here" or "in Paris". The utterance expressing this thought contains a linguistic expression (an indexical or a name) for a place.

When, again being in Paris, I think to myself or utter the true sentence "it's raining", my thought or utterance does not contain any constituent for a place. It does not make much sense, I think, to assume that "rain" is an ambiguous predicate, which sometimes expresses a meteorological relation between a time and a place and sometimes expresses a property of times. Nor, I think, is it plausible to assume that the fact that it's raining somewhere or other can make my weather-thought true. The weather-thought I would express by uttering "it's raining" in Paris is not ipso facto made true by the fact that it's raining in New York then. While in Paris, I may entertain a thought expressible by "it's raining" and grab my umbrella. If at this instant it's raining in New York and not in Paris, the right thing to say, I believe, would be that my thought is false, not true. In other words, although my weather-thought does not contain any constituent for a place, nonetheless it is not a general thought with an existentially quantifiable variable ranging over places. The problem then is to

capture the semantic relation between the place where rain is relevant to the truth-condition of my weather-thought and my thought, given that the latter contains no constituent for the former.

On Perry's view, in the above example, Paris (i.e., the place where I entertain my weather-thought) is an *unarticulated* constituent constituent of my thought. Although no place is articulated (or made explicit) in my weather-thought (or in my utterance), my thought is true iff it is raining in Paris, not anywere else. In a different example, a different place might be relevant. Suppose I'm in Paris speaking on the phone to someone in New York. Then by uttering "It's raining", it's conceivable that rain in New York, not in Paris, is relevant to the truth of what I'm saying and thinking. The unarticulated constituent of my weather-thought might therefore be New York. If, however, while in Paris, my weather-thought causes me to pick up my umbrella, then the unarticulated constituent of my thought is Paris. Perry's suggestion then is that my thought is not about Paris. Only thoughts including a constituent for a place can be about it. If Paris is not explicitly represented in my thought (any more than Z-land is ever represented in the weather-thoughts of Perry's celebrated Z-landers), then my thought concerns Paris. It's not about it.

Now, when I pretend that Sherlock Holmes was a smart pretentious detective who lived 221 Baker street in London, my thought concerns Conan Doyle's work. But it is not about it. I may engage in such pretence without representing explicitly the fact that my thought is about a piece of fiction. When I entertain the thought that many of Conan Doyle's works depict Sherlock Holmes as a smart pretentious detective who lived 221 Baker street in London, my thought is about Conan Doyle's work and hence is about the actual world. Similarly, when I entertain the pretend thought that witches have magical powers, my thought is about some non-actual state of affairs. It may concern Maria's beliefs, but it is not about them. When I entertain the thought that Maria believes that witches have magical powers, my thought is about Maria's beliefs, and hence about the actual world.

Finally, I want to mention one last point about simulation and cognitive architecture. I just mentioned difficulties for the view that thinking about beliefs is just a matter of simulating or pretending. I offered reasons for rejecting this view. On my view, thinking about beliefs is more complex than engaging in pretence: the former, unlike the latter, requires the deployment of the concept of belief. If so, then the ability to engage in pretence cannot be a sufficient condition for entertaining thoughts about thoughts or beliefs. This, I would argue, does require metarepresentational capacities – the ability to form representations about representations – not merely pretence. I now want to tentatively

argue that the ability to engage in pretence is not a necessary condition for being able to think about possible non-actual states of affairs either.

One might with good reason hold the view that engaging in pretence, simulating or imagining is more complex than merely believing something. It seems natural to assume that a creature may have beliefs (about her environment) without having the resources to engage in pretence. Plainly for a creature with linguistic abilities, making a pretend assertion must be something more complex than making a full blown assertion: the former presupposes the latter, but the converse does not hold. We cannot even begin to understand what it is to make a pretend assertion until we understand what it is to make an assertion. Nor can we begin to understand what it is to entertain a pretend thought or belief until we understand what it is to entertain a thought or hold a belief. Now the sense in which entertaining a bona fide belief is simpler than entertaining a pretend thought or belief is presumably that entertaining a thought or belief may be – although it need not be – about some actual state of affairs. Presumably, a creature must have the ability to form beliefs about some actual states of affairs (e.g., as a result of perception and/or memory) even though she does not have the ability to engage in pretence.

Now, it would, I think, be a mistake to argue – as Heal (1995) seems tempted to – from this difference in complexity between believing and pretending to the conclusion that the ability to engage in pretence must be a necessary condition for being able to entertain thoughts about possible non-actual states of affairs. The reason why it would be a mistake is that any creature capable of forming beliefs must also be capable of forming desires or having goals (or such motivational states). It must because unless it entertains desires or goals, it will not be able to act on the basis of its beliefs alone. Now, having a goal or forming a desire is representing a non-actual possible or impossible state of affairs. If so, then the ability to entertain goals or desires must involve the ability to represent possible (or even impossible) non-actual states of affairs. So pretence (or simulation) cannot be a necessary condition for entertaining representations about possible (or impossible) non-actual states of affairs. It cannot unless one defines pretence or simulation as the basic ability to form goals or desires.

For two complementary reasons, this, I think, would be unadvisable. Firstly, although one may disagree about its depth and its extent, there is some contrast between motivational states and informational states. The former function as goals, the latter as data-structures. Now it is constitutive of the former that they are about non-actual states of affairs. It is constitutive of the latter both they can be about actual states of affairs and that they can misrep-

resent them (or that they can be false). If pretence were by stipulation defined as the ability to form goals or desires (on the ground that they are of necessity about non-actual states of affairs), then entertaining a goal or a desire, unlike entertaining a belief about an actual state of affairs, would be entertaining a pretend thought. On this view, only beliefs about actual states of affairs would be full-blown thoughts. On the assumption that pretend thoughts are more complex than genuine thoughts,[10] this would imply that goals or desires are more complex than beliefs. There is, however, a sense in which desires must be simpler states than beliefs: this is the sense in which desires may be non-propositional (or non fully representational); in this sense, a desire may be directed towards some perceived object. According to some theory-theorists (e.g., Gopnik & Wellman 1994), human children seem able to master the concept of a non-propositional desire before they are able to pass the false belief task. This evidence, it seems to me, does support the view that desires are simpler than beliefs. Correspondingly, on the assumption that pretend thoughts are more complex than genuine thoughts, this evidence casts doubt on the view that entertaining a goal or a desire is entertaining a pretend thought.

Secondly, a creature with the ability to think about another creature's thoughts and/or her own must be able to distinguish her own desires or goals from the goals or desires of another. According to the simulation theory, I entertain somebody else's goals or desires as pretend-goals or pretend-desires. From my own standpoint, my own goals and desires are motivational states. Somebody else's desires or goals are not. At least, I don't act on the basis on somebody else's goals or desires. My own actions may indeed depend on my beliefs about somebody else's goals or desires. However, somebody else's desires are not my own desires; they may be pretend-desires of mine which I store in my belief box for the purpose of predicting somebody else's decisions. But I don't store them in my desire box. Of course, I may have higher-order desires about somebody else's desires as well as about my own first-order desires. Higher-order desires are genuine desires, not beliefs. But they are higher-order desires. They are not pretend-desires. Nor are they directly motivational as my own first-order desires are. So I don't think that the view that pretence or simulation is necessary for entertaining possible states of affairs has much to recommend itself.

5.2 Mimicry and the attribution of belief

I now come to a critical examination of the role of the fourth feature of the mental simulation heuristic: mental mimicry. Many simulation theorists de-

rive support for the mental simulation heuristic from the fact that mimicry is so widespread in human cognition. And so they should. Indeed, according to the mental simulation heuristic, when I am trying to figure out the content of another person's mental state, I produce in my mind a *copy* of the target state. Although I agree with simulation theorists that mimicry is a crucial element of human cognition, I nonetheless think they are wrong to assimilate mindreading and mental mimicry. First, I want to explain why I think the prevalence of mimicry in human cognition is an important finding from the standpoint of third-person mindreading. Second, I will explain why I think third-person mindreading cannot be a version of mimicry.

The goal of mindreading is to form an accurate picture of what our conspecifics think and have in mind. As theory-theorists have, I think, rightly emphasized, what other people do might be observable – their bodily motions are. But what they have in mind – their beliefs and desires – is not directly observable in their actions or behavior. Whether or not the mental simulation (pretence) heuristic is a fundamental part of our method for accomplishing third-person mindreading, our success in the task does presuppose a certain amount of psychological similarity between members of the human species. As simulation theorists have emphasized, existing mechanisms of mimicry will serve a useful purpose: they will enhance the probability of success in tasks of understanding the minds of one's conspecifics. For reasons that will emerge soon, mimicry, however, cannot be the process of third-person mindreading. Rather, it is an enabling condition for the accuracy of third-person mindreading.

Why cannot mindreading be mental mimicry? Emotional contagion, laughing, yawning, infants' copies of facial features and the firing of mirror neurons that underlie the replication of plans for specific goal-directed actions across individuals are all well-documented cases of mimicry. In all of these phenomena, the observation by individual B of some action performed by individual A triggers automatically in B a copy of either something A did or of some internal mechanism in A necessary for performing the action. There is a direct causal path leading from A, the agent, to B, the observer. But the very idea of copying is simply inapplicable to mindreading. When I try to understand one of your mental states, e.g., one of your beliefs, I try to capture two features of your mental state: its content and the attitude under which you entertain the content. The only way I can capture the attitude is by applying a mental concept (i.e., the concept of belief). The only way I can capture the content is by exploiting a representation whose content will be suitably related to the content of your belief.

Genes can be copied because, unless they mutate, they replicate. Utterances of sentences of natural and other languages can be copied or reproduced in virtue of their phonological non-semantic properties. Arguably, direct quotation and photocopying are methods for reproducing utterances. But when it comes to an appreciation of the content or semantic properties of utterances, reproduction or copying is not an option anymore. Only interpretation is available. Interpretation can occur across languages or within one language. Unlike direct quotation, indirect quotation is not a mere reproduction. Nor is, of course, translation, which consists in finding a sentence *s* in language *L* that is synonymous with (which has the same meaning as) sentence *s'* from language *L'*. Summarizing or exegetical expansion of an utterance, which take place within one language, involve finding other sentences in one and the same language to help determine the content of an initial utterance. The contents of two or more representations can stand in a similarity (or resemblance) relation of which the identity relation is, at best, an extreme limiting case.[11]

What is true of the contents of linguistic representations is true *mutatis mutandis* of the contents of mental representations. Unlike linguistic representations, mental representations lack phonological properties. Arguably though, assuming, as I do, that mental representations have neurological properties, techniques of brain imaging, might, unlike ordinary mindreading, allow one to reproduce a mental representation in virtue of copying some of its neurological properties. However, this very process in and of itself does not display the meaning or content of the copied mental representation. No more so, I submit, than photocopying a linguistic formula e.g., Quine's enigmatic slogan "To be is to be the value of a variable", in and of itself, is a method for determining its semantic property. Interpretation, not copying, is the way to get at the content or meaning of any representation.

In this paper, I have examined evidence for the role of mental simulation in the preparation of one's own actions and the visual detection of actions performed by others. Then I have examined the potential role of mental simulation in the mastery of such mental concepts as the concept of belief. First-person mindreading, I have argued, paradoxically constitutes a difficult challenge for the introspectionist version of simulation. There is a paradox here since the introspectionist version of simulation is based on the assumption that third-person mindreading depends on first-person mindreading. Finally, I have argued that not only does the anti-introspectionist version of simulation account wrongly identify ascribing a belief with engaging in pretence but also that representing others' beliefs is not mental mimicry.

Notes

1. I will freely use "simulation", "pretence" and "imagination" interchangeably.

2. For an extended justification of this rough classification, see Meini (1999).

3. As Jérôme Dokic has emphasized in his sympathetic and interesting comments, in a simulation process, pretence can occur on the input side, on the output side, or both.

4. For a review, see Jeannerod (1997).

5. *Ibid.*

6. Cf. Milner and Goodale (1995), Jeannerod (1994, 1997).

7. Some such view is argued for in Elster (1999).

8. For a case of anosognosia, see e.g., Bisiach (1988).

9. I shall question this assumption in the next section.

10. An assumption I accept.

11. This is a point made several years ago in Sperber and Wilson (1986). Cf. Jacob (1987).

References

Baron-Cohen, S. (1994). How to Build a Baby that Can Read Minds: Cognitive Mechanisms in Mindreading. *Cahiers de psychologie cognitive, 13*, 513–532.

Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (Eds.). (1993). *Understanding Other Minds. Perspectives from Autism*. Oxford University Press.

Bisiach, E. (1988). Language Without Thought. In L. Weiskrantz (Ed.), *Thought Without Language*. Oxford: Clarendon Press.

Campbell, J. (1994). *Past, Space and Self*. Cambridge, MA: MIT Press.

Carey, S. & Spelke, E. (1994). Domain Specific Knowledge and Conceptual Change. In Hirschfeld & Gelman (Eds.), *Mapping the Mind, Domain Specificity in Cognition and Culture*. Cambridge University Press.

Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon Books.

Currie, G. (1995). Visual Imagery as the Simulation of Vision. *Mind and Language, 10*, 1/2, 25–44.

Damasio, A. (1995). *Descartes' Error*. New York: Papermac.

Davies, M. & Stone, T. (Eds.). (1995a). *Folk Psychology*. Oxford: Blackwell.

Davies, M. & Stone, T. (Eds.). (1995b). *Mental Simulation*. Oxford: Blackwell.

Elster, J. (1999). *Strong Feelings, Emotion, Addiction and Human Behavior*. Cambridge, MA: MIT Press.

Fuller, G. (1995). Simulation and Psychological Concepts. In M. Davies & T. Stone (Eds.), *Mental Simulation*. Oxford: Blackwell.

Gallese, V. (2000). *Action, Goals, and their Role in Intersubjectivity: from Mirror Neurons to the 'Shared Manifold' Hypothesis. Mimeo*.

Gallese, V. & Goldman, A. (1998). Mirror Neurons and the Simulation Theory of Mind Reading. *Trends in Cognitive Sciences, 2*, (12) 493–501.

Goldman, A. (1989). Interpretation Psychologized. In M. Davies & T. Stone (Eds.) (1995a), *Folk Psychology*. Oxford: Blackwell.

Goldman, A. (1993). The Study of Folk Psychology. *Behavioral and Brain Sciences, 16*, 15–28.

Goldman, A. (2000). The Mentalizing Folk. In D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press.

Gopnik, A. & Meltzoff, A. (1997). *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

Gopnik, A. & Wellman, H. M. (1992). Why the Child's Theory of Mind Really *is* a Theory. In Davies & Stone (Eds.), *Mental Simulation*. Oxford: Blackwell.

Gopnik, A. & Wellman, H. M. (1994). The Theory Theory. In L. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind, Domain Specificity in Cognition and Culture*. Cambridge University Press.

Gordon, R. M. (1986). Folk Psychology as Simulation. In M. Davies & T. Stone (Eds.), *Folk Psychology*. Oxford: Blackwell.

Gordon, R. M. (1996). 'Radical' Simulationism. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Harris, P. (1992). From Simulation to Folk Psychology: The Case for Development. In Davies & Stone (Eds.), *Folk Psychology*. Oxford: Blackwell.

Heal, J. (1986). Replication and Functionalism. In M. Davies & T. Stone (Eds.), *Mental Simulation*. Oxford: Blackwell.

Heal, J. (1995). How to Think About Thinking. In M. Davies & T. Stone (Eds.), *Mental Simulation*. Oxford: Blackwell.

Hirschfeld, L. A. & Gelman, S. A. (Eds.). (1994). *Mapping the Mind, Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Jacob, P. (1987). Thoughts and Belief-Ascriptions. *Mind and Language, 2*, (4) 301–325.

Jeannerod, M. (1994). The Representing Brain: Neural Correlates of Motor Intention and Imagery. *Behavioral and Brain Sciences, 17* (2), 187–245.

Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Oxford: Blackwell.

Kahneman, D. & Tversky, A. (1979). The Simulation Heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainy: Heuristics and Biases*. Cambridge: Cambridge University Press.

Leslie, A. M. (1994a). *Pretending* and *Believing*: Issues in the Theory of ToMM. *Cognition, 50*, 211–238.

Leslie, A. M. (1994b). ToMM, ToBy, and Agency: Core Architecture and Domain Specificity. In L. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind, Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Leslie, A. M. (1995). A Theory of Agency. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition, a Multidisciplinary Debate*. Oxford: Oxford University Press.

Leslie, A. M. & Thaiss, L. (1992). Domain Specificity in Conceptual Development: Neuropsychological Evidence from Autism. *Cognition, 43*, 225–251.

Meini, C. (1999). La psychologie naïve, essai sur la nature, la fonction et les origines d'un dispositif cognitif spécialisé. Unpublished doctoral dissertation, Ecole Polytechnique.

Meltzoff, A. & Moore, M. (1983). Newborn Infants Imitate Adult Facial Gestures. *Child Development, 54*, 702–709.

Milner, D. & Goodale, M. A. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Perry, J. (1986). Thought without Representation. In J. Perry (1993), *The Essential Indexical*. Oxford: Oxford University Press.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Recanati, F. (1997). The Dynamics of Situations. *The European Review of Philosophy, 2*, 41–75.

Recanati, F. (2000). The Iconicity of Metarepresentations. In D. Sperber (Ed.), *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1995). Premotor Cortex and the Recognition of Motor Actions. *Cognitive Brain Research*.

Spelke, E. (1988). The Origins of Physical Knowledge. In L. Weiskrantz (Ed.), *Thought without Language*. Oxford: Oxford University Press.

Spelke, E. (1991). Physical Knoweldge in Infancy: Reflections on Piaget's Theory. In S. Carey & R. Gelman (Eds.), *The Epigenesis of Mind*. New York: Lawrence Erlbaum.

Spelke, E. et al. (1995). Infants' Knowledge of Object Motion and Human Action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition, a Multidisciplinary Debate*. Oxford: Oxford University Press.

Sperber, D., Premack, D., & Premack, A. J. (Eds.). (1995). *Causal Cognition, a Multidisciplinary Debate*. Oxford: Oxford University Press.

Sperber, D. & Wilson, D. (1986). *Relevance, Communication and Cognition*. Harvard: Harvard University Press.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Stich, S. & Nichols, S. (1992). Simulation or Tacit Theory? In M. Davies & T. Stone (Eds.), *Folk Psychology*. Oxford: Blackwell.

Stich, S. & Nichols, S. (1995). Second Thoughts on Simulation. In M. Davies & T. Stone (Eds.), *Mental Simulation*. Oxford: Blackwell.

Ungerleider, I. G. & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of Visual Behavior*. Cambridge, MA: MIT Press.

Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Reply to Pierre Jacob

Jérôme Dokic
Université de Rouen and Institut Jean-Nicod, Paris

Let me say at the outset that I am in basic agreement with what Pierre Jacob says in his paper. What I shall do here is, briefly, expanding on one argument he gives in Section 5 of his paper against radical simulationism as defended by Robert Gordon. Then I shall mount another, perhaps more tentative argument against this version of simulation theory. Finally, I propose to qualify Jacob's stronger claim that pretence or simulation is not even necessary for having beliefs and desires.

I. Gordon defends what might be called the *non-conceptual view of belief-ascription*, according to which the capacity to ascribe beliefs does not require the antecedent possession of the concept of belief as a mental state. Thus, the ability to express a worldview within the scope of simulation can be the most central part of the capacity to ascribe beliefs to the person holding that worldview. If some cases of simulation can ground belief-ascriptions, and do not require mastery of the concept of belief, belief-ascriptions themselves do not have to deploy such a concept.[1]

In contrast, Jacob rejects the claim that to make a belief-ascription is to make an assertion, or to express a belief, in the context of practical simulation. One of Jacob's objections to Gordon's account is that belief-ascriptions (when true) state mental *facts* about the actual world, whereas simulation can only yield *pretend* mental states. More generally, it is one thing to *engage* in simulation, but it is another thing to *describe* or *exploit*, in a detached way, the process and results of simulation. The problem Jacob sees in Gordon's account is that a simulation process culminating in a mere pretend assertion falls considerably short of grounding a belief-ascription. Using a distinction drawn by John Perry, a belief-ascription is *about* someone's belief, whereas a simulation at best

concerns that belief. The gap between the former and the latter must surely be extensive, as is the one between Z-landers and us.[2]

Now, of course, we have to recognize the distinction between free imagination and what we might call *anchored simulation*. When I pretend to be in Mary's situation, my simulation is *about* Mary; it does not merely *concern* her (accidentally, so to speak). How is it possible? First, as Gordon says, we try "to make adjustments for relevant differences", i.e. we feed the simulation with inputs relevant to Mary's situation. I shall assume that there are basic cases of simulation in which these inputs are not mental states; otherwise the very process of anchored simulation would presuppose the antecedent ascription of such states to Mary. For Gordon's purposes, it is of course crucial that the inputs are taken from Mary's behavioral and situational evidence and do not involve mental states as such (but rather postures, gestures, lines of gaze, observable interactions with objects, and so on).

So simulation is anchored because it takes into account the simulated person at the input side. However, it might be argued that it *must* be anchored *at the output side* as well. Jacob raises the problem of what *use* we are to make of the results of simulation. Fuller has expressed a similar worry in the following way:

> [W]hat is the ascriber supposed to *do* with the pretend assertion? Surely, if any pretend assertion (or what it is about) is to be part of an account of what I mean when I say that S believes that *p*, then it will be related to S in some way. But in what way? (1995: 26)

The intuition behind this worry is that simulation can ground ascriptions of mental states only if it is *doubly anchored*, at the input side as well as at the output side, where its results are *used* by the subject in some way or other. An important way in which one might use the results of simulation (and thereby step outside the simulation process) is to modify or adjust one's own actual behavior, taking into account the other's moves (as they are predicted by the simulation).

Without doubt this is a perfectly legitimate use of simulation. Note, though, that it is entirely *practical*. All that it offers is a prediction that *given* some non-doxastic inputs (i.e. inputs other than beliefs or more generally mental states), other non-doxastic outputs are to be expected (or are judged to have taken place, if we simulate a past situation). Here, I can use the results of simulation without explicitly ascribing any particular beliefs, because simulation goes all the way through to pretend intentions or motor representations, which

can be *directly* taken into account in modifying my own behavior in response to yours as it is predicted by my simulation.

However, and precisely for this reason, it is difficult to see how a conception of *particular* beliefs could be extracted from this form of simulation. Given the use being made of the results of simulation, how is the subject supposed to grasp the essence of a belief state, which is, as Gareth Evans said, that of being "at the service of many distinct projects", and whose influence on any project is mediated by other beliefs (1985: 337)? Simulation is used here only for bridging the gap between (pretend) non-doxastic states, so how could a conception of intermediary mental states like belief emerge from it?

The friend of radical simulationism might reply that simulation does not always go all the way through to motor representations; it sometimes stops at, or "culminates" in a particular pretend-belief (or pretend-assertion). This raises Jacob's and Fuller's worry again: what does it mean to *use* the results of simulation when it stops at a particular pretend-belief? How am I supposed to move from this stopping-point to the adjustment of my behavior in response to yours? The point is not that it is impossible to use the results of *partial* simulation (simulation that stops at pretend doxastic states). We do this all the time. The point is rather that it is difficult to see how we can do this without deploying an antecedent conception of belief. I cannot deduce the way you are going to behave from a pretend-belief, since I have to make hypotheses about your *other* beliefs and mental states, and the way they combine to rationalize behavior. I cannot step outside the process of simulation "in between" unless I already have the theoretical means to relate the intermediary results to the non-mental world.

To sum up the first argument, non-conceptual simulation (i.e. simulation that does not require mastery of the concept of belief) makes sense only if the results of simulation can be *directly* used to control my behavior, without going through an independent process of belief-ascription. I have argued that such simulation is certainly useful, but cannot ground an understanding of the impact of a particular belief on behavior, because it cannot stop at a pretend belief but must go through to pretend states or processes that enable me to adjust my behavior in response to the other's. In a nutshell, Gordon's account faces the following dilemma. Either simulation stops at motor representations, but cannot be used to ground an understanding of the contribution of particular beliefs to behavior, or it stops at particular pretend beliefs, but cannot be anchored at the output side without an antecedent understanding of what a belief is.

II. What I have argued so far – I think in line with Jacob's paper – is that there is something like non-conceptual, practical simulation, but that there is a substantial gap between it and a more reflective, detached understanding of mentality involving explicit concepts of mental states such as belief. Now one possible line of retreat would be to *insist* that when simulation is non-conceptually anchored at the output side (i.e. goes all the way through to non-doxastic states like motor representations), the subject has implicitly ascribed all the intermediary beliefs to the simulatee. On this line, it is conceded that there is a gap between a practical form of belief-ascription based on simulation and a more detached form of belief-ascription which explicitly deploys the concept of belief. However, it is claimed that both forms of attributions ascribe the same kind of states, namely beliefs. The point is that they ascribe them in different – practical and detached – ways.[3]

What I want to do now is to present a tentative argument which casts doubt upon the legitimacy of this line of retreat. The conclusion of this argument is that *if* anchored simulation is conceived as a form of non-conceptual ascription of mental states, there are reasons to think that these mental states are *not* of the same kind as the ones we explicitly attribute using ordinary belief-ascriptions.

My argument starts with a reminder: belief is an attitude with a propositional content. To grasp the essence of belief as a propositional attitude, it is not enough to realize that any belief is true *or* false. One also has to understand that at least some beliefs (or some propositional contents) can be true *and* can be false.[4] Mastering the concept of belief as a propositional attitude requires that we understand the metaphysical possibility that some of our beliefs are false.[5]

Now it can be argued that the understanding of belief as something that can be true and can be false cannot be grounded on, or extracted from, simulation alone. My strategy in arguing for this claim will be to identify a conceptual scheme which is much less sophisticated than ours. A subject with such a conceptual scheme will constitutively tie "the facts" to the world as it presently appears to her. In particular, there is no way for her to envisage the possibility that her *present* worldview (as it is determined by her beliefs) is false (although I do not mean to suggest that such a conceptual scheme is actually realized in some, particularly stubborn creature).

It is worth pointing out that a subject having this conceptual scheme might pass at least some versions of the false-belief task. In other words, she might be able to acquire the notion of a representation that is at variance with "the facts". As Gordon himself says, the subject "is not yet in a position to understand that *her own present beliefs* may themselves deviate from the facts" (1995: 62). Now

what is remarkable is that Gordon thinks that a form of *embedded simulation* can yield the necessarily more sophisticated notion of belief:

> To see her own present beliefs as distinguishable from the facts she will have to simulate another from whom the facts are different – or, more broadly, adopt a perspective from which the facts are different, whether this perspective is occupied by a real person or not – and then, from the alien perspective, *simulate herself.* (1995: 62)

In fact, the ability to simulate oneself from another simulated perspective does not require departing from the simple conceptual scheme described above. It would require such a departure if the subject had the means of conceiving that the other *might be right* in thinking that one's own worldview is wrong. Embedded simulation can at best yield an understanding that others might think that one is wrong; it cannot yield an understanding that others might be *right* in thinking that one is wrong. It follows that it cannot yield an understanding of my beliefs as something that I take to be correct (by hypothesis), but which *can* be incorrect.

Gordon described another way in which, according to him, one may come to understand that one's present worldview can be wrong. It may arise with "the capacity to demote some of one's own *memories* to off-line status" (1995: 62). I seem to remember that this was a rock, but I now know that this is a sponge. I come to understand that my previous worldview was wrong. Unfortunately, the capacity to demote some of one's own memories to off-line status, as the capacity to simulate a possible later perspective of one's own, is insufficient to ground the crucial idea of a single belief capable of being true *and* capable of being false. I might be able to conceive of a later perspective according to which I am now wrong, but I might not understand that such a perspective could be the correct one. I might understand that I was wrong in the past, but I might be unable to understand that I could have been right.

My tentative conclusion is that simulation, however multiply embedded, cannot ground an understanding of belief as having a propositional content which can be true and can be false. My argument for this conclusion is that simulation is always compatible with a less demanding conceptual scheme, according to which what we deem as "the facts" is simply equated with our present doxastic world. An understanding of propositional attitudes requires much more than practical simulation. In my view, it requires at least a grasp of the subject-predicate structure, and a reflective conception of how a thought is determined by its position in a network of normative, functional relations essentially involving other thoughts.

If we accept the second argument, it cannot be maintained that practical simulation is enough by itself to ground implicit ascriptions of the mental states *we* call "beliefs". As the gap between practical simulation and detached understanding of mental states has got still more extensive, radical simulationism has been correspondingly weakened.

III. So I tend to agree with Jacob that the ability to simulate is not a sufficient condition for being able to ascribe beliefs to someone else or to oneself. However, Jacob also wants to defend a stronger claim, namely that the ability to engage in pretence is not even a necessary condition for being able to think about possible non-actual states of affairs. Simulation, pretence or imagination (Jacob makes clear that he uses these terms interchangeably) is not involved in the very process of representing possible non-actual states of affairs. I should like to qualify this stronger claim, in fact reject it if "representation" here means propositional representation, of something that can be true and can be false.

One argument Jacob gives in favor of this claim is that a creature may have beliefs about her environment without having the resources to engage in pretence. However, it seems to me that this argument works only if the notion of belief involved is that of a *primitive* belief, something that precisely does not have the property of bipolarity characteristic of most of our propositional beliefs. For a propositional belief depicts a possible state of affairs independently of whether that state is actual or not. Of course it is *believed* to be actual, but the subject must have the resources to understand that this very belief might be false. I claim that such capacity requires the imagination. One cannot grasp a bipolar content, and so one cannot have the corresponding bipolar belief, if one does not have the ability to imagine relevant possibilities.

Of course, as Jacob points out, even a primitive belief-system must be capable of misrepresenting. There is a distinction to be made, though, between misrepresentation and bipolarity. A belief-system can misrepresent if it sometimes produces non-veridical representations. It does not follow that these token representations are bipolar, i.e. can be veridical and can be incorrect. Among beliefs, only propositional (and so conceptual) ones have such a property. So I agree with Jacob that primitive beliefs can combine with non-propositional desires to produce intelligible behavior without the help of the imagination. What I should like to resist, though, is the conclusion that *our* propositional desires, although they are about non-actual, possible states of affairs, do not essentially involve simulation abilities.

Notes

1. Of course simulation *sometimes* involves beliefs about mental states, for instance if I simulate a piece of defeasible reasoning which, according to some epistemologists, requires my having beliefs about the reasoning itself. Gordon's point is that simulation *need* not involve such meta-beliefs.

2. Here I refer to the characters described by Perry (1993), who can think and say "It's raining" but are unable to think of other places where it might not rain.

3. Of course this line faces the problem of explaining how detached belief-ascription can emerge from practical simulation, given the gap that there is now between them.

4. Readers of Wittgenstein will recognise his early definition of a proposition, transposed here into an account of empirical beliefs.

5. Of course, we can be wrong in thinking that a particular belief of ours can be false in this sense. Perhaps the falsity of "Cats are animals" is an epistemic possibility, even though it is not a metaphysical possibility.

References

Davies, M. & Stone, T. (1995). *Mental Simulation*. Oxford: Blackwell.

Evans, G. (1985). *Collected Papers*. Oxford: OUP.

Gordon, R. (1995). Simulation Without Introspection or Inference from Me to You. In Davies & Stone (Eds.).

Fuller, G. (1995). Simulation and Psychological Concepts. In Davies & Stone (Eds.).

Perry, J. (1993). Thought Without Representation. In *The Problem of the Essential Indexical*. Oxford: OUP.

Chapter 5

Some reasons to link imitation and imitation recognition to theory of mind

Jacqueline Nadel
CNRS Unit Cognition and Communication, Paris

An interesting new idea that emerged in recent debates about how a Theory of Mind (TOM) develops in young children consists in recognizing that what goes on in minds is naturally of interest to humans, to put it in Trevarthen and colleagues' words (Trevarthen, Kokkinaki, & Fiamenghi 1999). But the problem of how could an intention be transferred from one person to another remains to be clarified. How could behaviors and their motives be intersubjectively transmitted? Which role do such transfers have in understanding our minds and those of others? The analysis of these transfers is of special interest to simulationism.

As a psychologist working in the realm of early development, I will focus on four aspects that may account for a precocious propensity to conceive persons as intentional, and may prepare a further capacity to simulate:

1. Neonates prefer and react preferentially to human stimuli.
2. Young infants detect contingently responsive behaviour.
3. Young infants expect contingency and thus are able to form ontological knowledge about persons as communicative agents.
4. Once they consider imitation as a contingent response, young children are ready to communicate via imitation.

I will explore these four points with a special focus on the role of imitation in early understanding of others as agents.

1. Neonatal preferences for human stimuli

a. Perception

Neonates differentiate human from physical stimuli: they prefer human odours, human faces, human sounds. For instance, a preference for maternal odour (Schaal 1988), and for maternal milk over other matched odours has been found (Soussignan, Schaal, Marlier, & Jiang 1997). With the habituation paradigm, newborns have been shown to differentiate human faces from scrambled faces. They do appreciate some configural properties of an adult's face at birth (see Muir & Nadel 1998, for a review). Newborns' preference for the mother's voice (DeCasper & Fifer 1980), as well as for maternal language (Mehler et al. 1988) has also been demonstrated. While newborns pay so much attention to external stimuli, and especially to human stimuli, as part of their prior history of environmental stimulation, they do not seem to make differential social responses at birth, except imitation of a few facial movements.

b. Response

Neonates respond to human stimuli via imitation. They do not need to learn how to imitate. After the seminal studies by Zazzo (1957), Maratos (1973), and Meltzoff and Moore (1983), number of studies have documented neonatal imitation of tongue protrusion and mouth opening. Other studies have shown that neonates can also imitate eye blinking (Kugiumutzakis 1993). They do not confuse the facial part in play: they blink eyes in response to eye blinking, not to mouth opening, they protrude tongue in response to tongue protrusion, not to mouth opening. They only imitate human movements. To our knowledge, only one study (Jacobson 1978) accounts for infants' imitation of an object simulating tongue protrusion (a pen rhythmically going back and forth) and mouth opening (a ball rhythmically going up and down), but the infants were 6 to 14 week-olds, and Jacobson's results have never been replicated.

When they imitate, newborns not only demonstrate a selective attention to human stimuli and an early motor capacity, they also seem to give a social response. Indeed, even if they do not plan to do so, they produce temporal contingency since they imitate a movement just after it has been performed; they also produce morphological contingency since their movement is similar to the one displayed by the adult. The enthusiastic reaction of parents when they see their infant imitating a tongue protrusion is a very convincing index of

the fact that imitation is probably the unique signal of social contingency available for parents at birth, while also an infrequent phenomenon in naturalistic conditions (Uzgiris, Vase, & Benson 1984).

2. Young infants' capacities to detect social contingency

Although imitation seems to be the first contingent answer produced by infants facing human stimuli, we have no reliable indices in favour of the idea that infants detect imitation prior to any other contingent behavior.

Indeed, after a few weeks, young infants engage selectively with persons who have an upright face (not an inverted face) and only one head, who smile, who make eye-to-eye contact (they detect a 20 degree turn away and seem to dislike it), who present facial movements, and display contingent interaction. Whether this preference only accounts for a distinction between dynamic and static devices, or reveals an innate capacity to understand intentionality is still hotly debated. For instance, Gergely and Watson (1998) postulate the presence of an innate module that detects perceptual contingency. According to this view, there is a common cognitive origin to both physical and social awareness and control of contingency. However, some studies failed to find relationships between physical and social targets, which calls for an examination of the affective component of awareness and control of contingency by infants (Snow 1998). An alternative hypothesis is that detection of social contingency, social responses to contingency and expectancies for social contingency may be three relatively independent components of one and the same innate capacity to understand intentionality and to process agency. Within this framework, the early existence of expectancies for social contingency by infants could be considered an index of the ontological understanding of human behavior as intentionally social.

To document the hypothesis that human infants read goal-directed mental states into a wide range of stimuli having inherent direction of self-propulsion, researchers designed experimental paradigms that experimentally violate social rules of positive communication. Such paradigms allow testing whether or not young infants have already formed expectancies about their close partners' social behaviors (Stern 1977; Murray & Trevarthen 1985; Tronick 1977; Muir et al. 1995; Nadel et al. 1999). Among the handful of paradigms available, the more powerful are the still-face design and the double video live-replay design. The still-face design is based on an unexpected change in the mother's behavior, suddenly becoming inexpressive, still and silent. Numerous studies

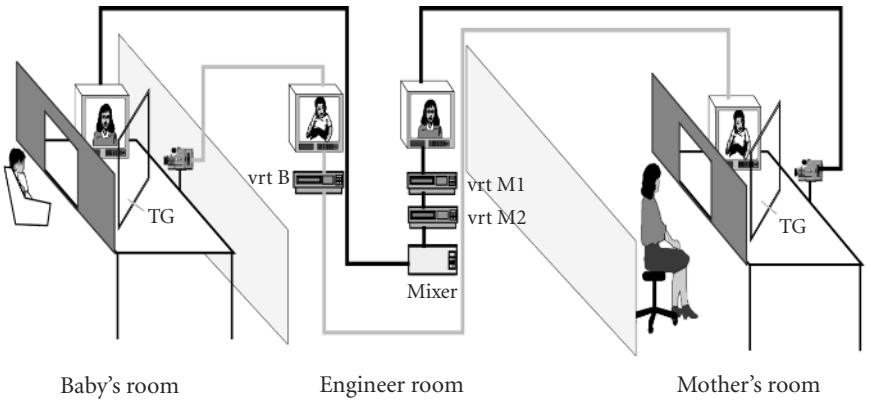

Figure 1. Experimental device[1]

(Field 1977; Fogel et al. 1982; Murray & Trevarthen 1985; Tronick et al. 1978), demonstrated that the infants gazed dramatically less to mother and clearly displayed strong emotional disturbances during the still episode (see Muir & Nadel 1998, for a review). The emotional disturbances concerned persons but not dynamic objects such as for instance a contingent doll, thus indicating a distinctive reaction to still objects versus still persons (Ellsworth et al. 1993; Legerstee et al. 1987).

The double video live-replay design, created by Murray and Trevarthen (1985), exposes the infant to more subtle violations of social communication, such as violations of conversational synchrony. The infant's social behavior is compared in two conditions of televised interaction: a live face-to-face with the mother through video-monitors, and a replay of mother's previous communicative behavior. Is a young infant sensitive to the mismatch between her behavior and her mother's response?

To answer this question, we modified the design in order to have a three-condition experiment (first live-replay-second live) with seamless shifts from live to replay and from replay to live, so that the infant could continuously see and hear her mother, either live or replayed. We proposed to ten 9 week-olds and their mothers to interact live through a closed circuit of TV monitors. After a period of good contact with their mothers, the infants were unexpectedly presented with a good communicative replay of maternal behavior during 30 seconds, and were then again unexpectedly allowed to interact live with their mothers (see Figure 1).

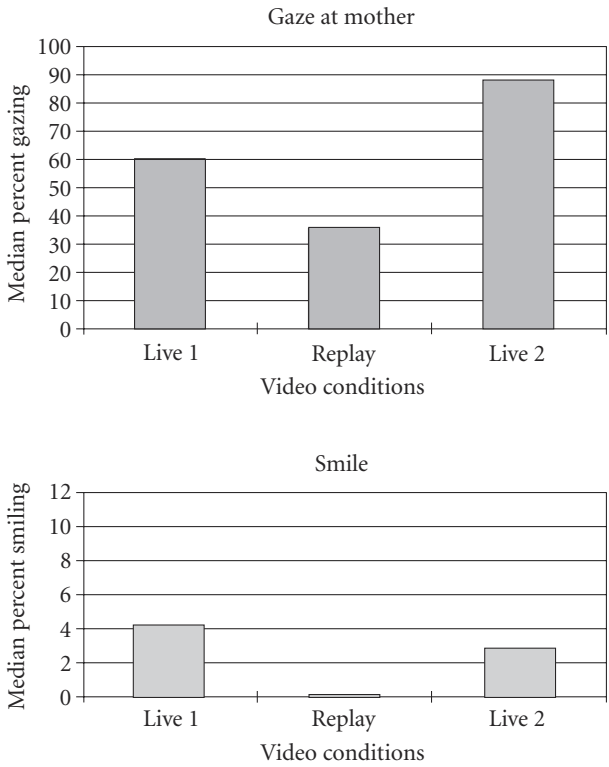

Figure 2. Positive behaviors accross conditions

Although the procedure was different from the procedure used by Murray and Trevarthen (1985) and the detection of a change in interaction was more difficult due to the seamless shift from live to replay, our results clearly showed that infants as young as 2-month-olds were able to detect non-contingency. They turned away from mother and withdrew during the replay episode and interacted again positively during the second live episode (see Figure 2).

We found an inverted U-shape for negative behaviors (frowning and mouth closed). These behaviors increased during non-contingent interaction and decreased during the second live episode (Figure 3).

Now let us turn back to the experimental procedure and remember that during the non-contingent episode the infants saw their mothers gazing at them and smiling to them, and heard their mothers talking to them kindly. Is this an upsetting situation for the infants? Only of course if the 2-month-olds have already formed an implicit understanding that their close partners are not

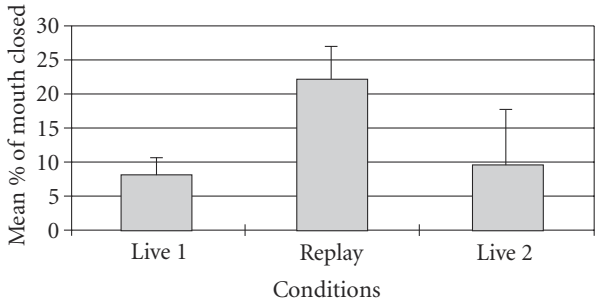

Figure 3. Month closed across conditions

only animate but also intentional beings responding or not responding according to their willing. To us, the behavioral changes during the non-contingent episode reflect a fundamental aspect of early intentionality detection. This implicit awareness of the partner as an intentional agent is however restricted to the cases of prior exposure to social contingent behavior of a close partner (Nadel et al. 1999).

3. Expectancies for social contingency and ontological knowledge of agency

A few weeks later, around 5/6 months, dyadic experimental paradigms allow researchers to test the infants' basic knowledge about the characteristic features of interaction with humans, compared to interaction with objects. For example, Hains and Muir (1996) demonstrated that 5 month-olds have formed a generalized expectancy that strangers will initially engage them in reciprocal interactions. When the stranger fails to do so in a non-contingent episode, visual attention and positive facial expression decreased. Similarly, Reyes et al. (1998) showed that 6-month-olds facing two strangers explore more the still stranger compared to the communicative stranger. These results account for precocious expectancies in infants for human beings – even strangers – to show contingent responsiveness. Six month-olds have already formed primary ontological expectancies about human behaviour.

In a pilot study, we tested whether low-functioning children with autism can also form a generalized expectancy about the social behavior of a stranger. To this end, we designed a three minutes-three conditions procedure composed

Table 1. Procedural stages of the revised Still Face paradigm

	Still Face 1 SF1	**Imitation Interaction II**	**Still Face 2 SF2**
Duration	3 minutes	3 minutes	3 minutes
Description	Child enters alone a familiar room, where an unfamiliar adult is sitting, still face and still body, silent, looking at the window	Experimenter stands up, walks to the child, smiling, and imitates all what the child does with the similar object	Experimenter sits again on the sofa, still face and still body, silent, looking at the window

of a Still Face episode, followed by an interactive episode, followed by a second Still Face episode, as summarized in Table 1.

In contrast to MA matched control children (MA = CA = 30 months) who often left the room, cried or hovered near the door, none of the autistic children showed concern with the still behavior exhibited by the stranger during the first still face episode. The important finding is that all the children with autism showed concern when the stranger exhibited again a still face after an imitative interaction episode (see Figure 4). In sum, they reacted to social rule violation after the adult had given proof of her humankind, but were not able to form social expectancies independent of prior experience (Nadel et al. 2000).

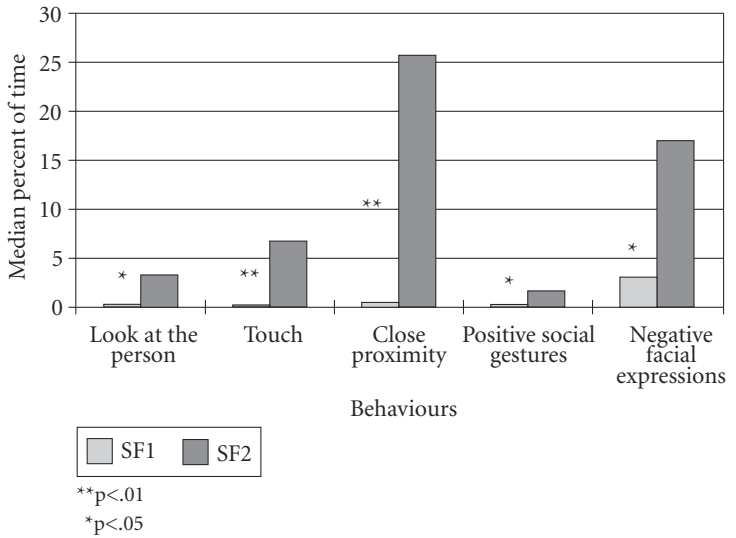

Figure 4. Significant changes in social behavior across Still Face conditions[2]

Dyadic paradigms such as live-replay TV and still face designs are powerful tools to test social perception and basic knowledge about the characteristic features of interaction with humans. Results of very young infants and of children with autism suggest a two-stage access to implicit understanding of intentional contingency in social interaction: first restricted to prior exposure to a contingent behavior of the partner, expectancies will later be generalized to any new human partner.

Sensitivity to contingency and sensitivity to imitation

We have drawn the picture of a very young infant, under 6 months, who is able not only to detect non-contingency but also to expect contingent behavior from close partners, and who starts forming generalized expectancies concerning human beings (in contrast with low-functioning children with autism). What can we say about the infant's ability to detect and expect imitative behaviors during this period? Since this special category of contingent behaviors occurs not only "*just after you*",[3] like any contingent behavior, but also "*just like you*",[4] they should be also detected and expected very early. If we follow Meltzoff and Gopnik (1993), they are detected even earlier than non-imitative contingent behavior.

The like-me mechanism

According to Meltzoff and Gopnik (1993), the neonatal capacity to imitate accounts for an interpersonal coupling which may well form the innate basis, the "starting-state mechanism" for the later elaboration of common-sense psychology and TOM. Indeed Meltzoff and Gopnik postulate a *like-me mechanism*, which provides a link between the self and other: the infant is able to produce a seen act and to recognize an imitated act like a *like-me* act. Human acts are especially relevant to infants, they say, because they look like the infants feel themselves to be; and also because human actions are things that the infant can intend to perform. This, however, remains to be demonstrated. Although it is now well established that neonates are imitators, it does not follow that they are also imitation recognizers from start. Indeed it is not granted that the detection of an equivalence between what we do and what we see requires the same capacity as the detection of an equivalence between what we see and what we do. Against this postulate, we could argue that early imitation (and some other kinds of direct imitations) may be explained as simple perceptual-motor coupling, an explanation which cannot account for imitation-recognition. Second, imitation may unintentionally convey social contingency, while imitation recognition implies attributing intentionality or agency to the imitator. This

would lead me to predict a later access to imitation recognition than to imitation. And as a matter of fact, while the ability to imitate is so richly documented, we have no empirical demonstration of a neonatal ability to recognize being imitated.

To document this question, we used our TV face to face design with 20 two month-olds, observed some maternal imitations but did not find any evidence at all of a sensitivity to being imitated. Similarly, Uzgiris et al. (1984) in their naturalistic follow-up of 3-to 12 month-olds interacting with mothers found maternal imitation of 3-month-olds in 20% of interactive episodes and only 12% of two-round episodes. In two-round episodes, either the infant, or more probably the mother, imitated in return. These data suggest that 3-month-olds' recognition of being imitated is at least very infrequent, if observable. With older infants, after 7 months, Field (1977) found more gaze to mother when the mother imitated her 32 week-old than when she interacted in another way. Does this account for imitation recognition? It is difficult to answer this question, since the author herself indicates that the mothers behaved differently in several ways while imitating their children. In any case, the infant's sensitivity to non-imitative contingency looks more precocious than the infant's sensitivity to imitative contingency. How can this been explained? First, and from a methodological point of view, experiments aimed at demonstrating infants' sensitivity to contingency in fact demonstrate sensitivity to non-contingency, which may be easier to demonstrate than sensitivity to contingency. Second, it is possible, as suggested by Meltzoff and Moore (1999), that there is a precocious implicit awareness of being imitated. Such awareness is supposed to be captured by the increase of attention to the imitator. It remains however that to make sure that a child attributes agency to an imitator, clear behavioral indices, such as strategies to test the imitator, are required (after all, 3 month-olds use showing off strategies to test their still mother).

Using a cross-target method, Meltzoff (1990) demonstrated that 14 month-olds prefer adult imitative contingency to adult non-imitative contingency. The contingent imitator received more gaze and smile, and more testing behaviors aimed at checking whether the adult was intentionally imitating them. This accounts for imitation detection and attribution of agency. As Meltzoff found this pattern down to 9 months, I will conclude that to-date, the first clear evidence of imitation recognition appears not earlier than 9 months, and under experimental conditions which may increase the infant's attention to the imitator. In more naturalistic conditions, most but not all 18 month-olds showed awareness of being imitated by another toddler, as measured by imitating the partner in return. Similarly, with 18- to 24-month-olds, several authors demonstrated

that being imitated prompts further imitation of the model. Thus clear attribution of agency to the imitator is not documented before 9 months in healthy children, long after these children started to imitate.

To document the capacity of imitation recognition in children with autism, we analysed the reactions of 25 children with autism in a situation where they were systematically imitated by an experimenter. All the children were occasional or frequent imitators. As an index of imitation recognition, we computed the number of switches between looking at the partner's hand or object, and looking at one's own object or hand. Data showed that 1/3 of the children did not look at the adult while she was imitating them emphatically, did not test her behavior at all, although they did not show any gaze aversion. There was no evidence that they were able to detect imitative behaviors (Nadel 2002).

If we consider that very young infants as well as some low-functioning children with autism can imitate but do not show observable indices of imitation recognition, it is difficult to hypothesise that the postulated *like-me mechanism* works bidirectionally at birth as an analyser of transmodal isomorphism from 'seen to done' and from 'done to seen being done'. At least, the two facets of imitation have unequal visibility. In order to recognize being imitated, we have to attribute intentionality to the partner. If we do not, the imitation will not appear to be such, or will be considered similar by chance. But now think of what happens if you are able to attribute to your imitator the intentional choice to imitate you. Your attention is attracted, you are self-confident, you know you have a partner interested with you and trying to communicate with you. As you are able to imitate too, you can reciprocate in the same way. It is exactly what preverbal children do after around 18 months of age and until the mastery of speech, around 42 months (Nadel & Baudonnière 1982; Nadel 1986; Nadel & Pezé 1993). Doing so, they use a communicative system that includes the two necessary and sufficient features of a language: production of social contingency (via imitation) and detection of social contingency (via detection of imitation).

4. Imitation and imitation detection, two facets of a communicative system, two facets of a computation of intentionality

When we started to hypothesise that imitation could be a basic language, we had two statements to demonstrate: (1) that imitation is a communicative system and (2) that imitation is a basic language. How can we show that imitation is a communicative system? It should present the criterial parameters of com-

munication: reciprocity (turn-taking), shared meaning (signals, indices, symbols), and long exchanges (more than 2 turns), and, above all, intentionality (monitoring and control). How can we show that imitation is a basic language? If it is available before speech, and if it disappears after age 3, when speech is mastered.

Our pilot study, conducted in 1979 explored the hypothesis that imitation is a communicative tool allowing to monitor long-lasting exchanges among preverbal children (Figure 5). Acquainted thirty month-olds met in triads without an adult present in a room filled with different kinds of attractive objects, some aimed at developing bodily postures (cow-boy hats, sunglasses, umbrellas), others at pretending (dolls, stuffed animals), others at demonstrating physical abilities (balls, inflated balloons, clinging mobiles). All the objects could be used solitarily, for their intrinsic value, or as mediators of social interest. A third possible option was to combine the two interests in using objects as social mediators.

Let us complete our description of the design by adding that three identical sets of all the objects were available and that the sessions were filmed with a hidden camera.

Results show that children chose most often to handle the object(s) similar to the object(s) hold by the other(s) at the same time: Simultaneous holds of identical objects were 3 times more numerous and longer than holds of different objects. Children were mainly partner-focused when they held identical object while they were half and half object or partner-focused when they held different objects. Ninety percent of imitative activities were developed during the hold of identical objects. We concluded that the hold of an identical object is an excellent index of immediate imitation (only 3 sequences of imitation with a different object were found out of more than 1500 sequences coded) and that imitation without object is very rare.

Another interesting result is the high positive correlation per child between the number of times they imitated and the number of times they were imitated. This significant correlation accounts for a symmetry between the two roles of imitator and imitatee; it is even more significant here with triads since there are only two roles for three persons.

Imitative sequences were initiated by an offering of the identical item or by choosing an item while looking at the partners and then waiting for them. The child who was proposed the identical item mostly took it and started imitating the initiator's activity. Sometimes the partner refused the initiator's suggestion. In this case the initiator left its item and turned to imitate the partner's ongoing activity. This accounts for the use of primary conventional rules to moni-

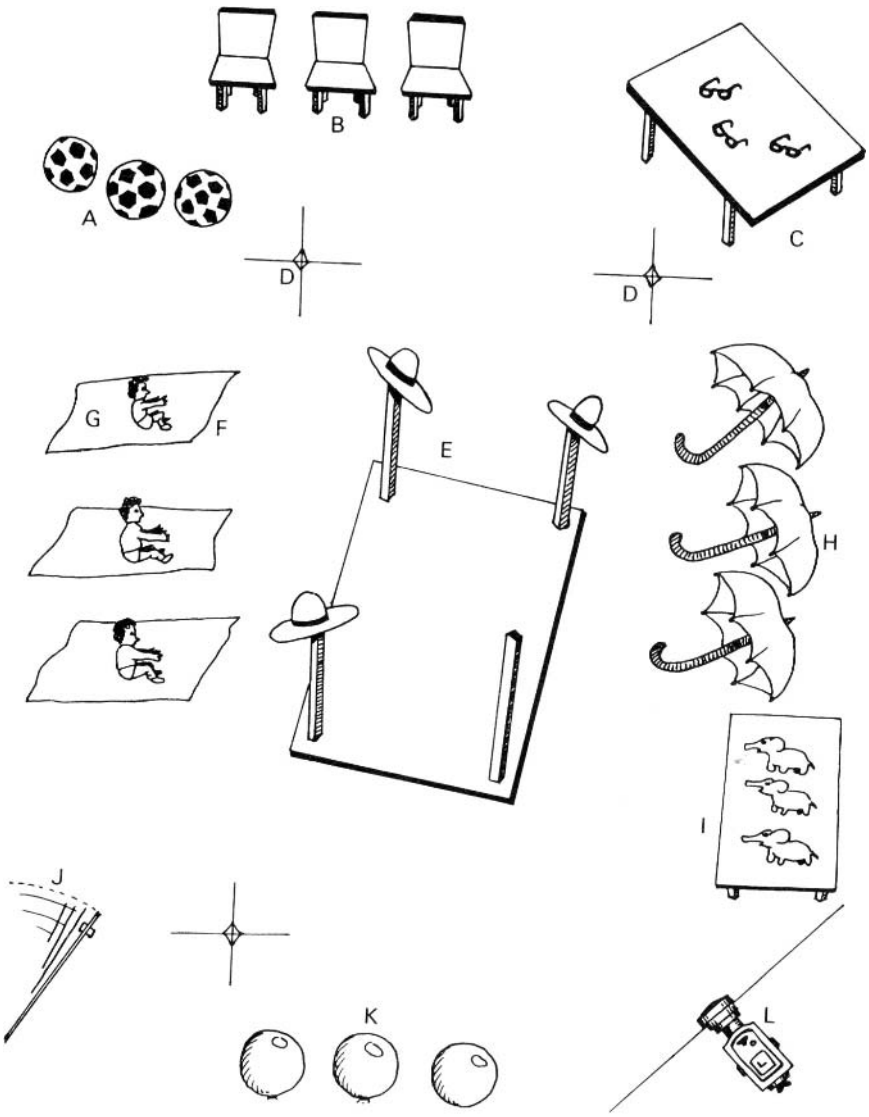

FIG. 1. — L'espace expérimental en condition à objets multiples

(A) ballons; (B) chaises; (C) lunettes sur table; (D) mobiles suspendus au plafond; (E) chapeaux de cow-boy sur table retournée; (F) serviettes bariolées avec poupées (G); (H) parapluies; (I) animaux en peluche sur table; (J) porte accessible aux enfants; (K) ballons de baudruche; (L) caméra vidéo mobile, avec zoom.

Figure 5. Imitation design

tor and control the other's activity, and to organize turn-taking, but also and principally for the computation of intentionality. Indeed, when initiating an imitative sequence, an infant has to work out whether or not, when he offers or points to the identical item, the partner(s) will be willing to imitate; in case he is not, he will have to switch roles and imitate in turn; the infant has to compute, when imitating, whether or not the imitatee detects the imitation, and accordingly to repeat the imitation or switch role and attempt to be imitated. Another computation related to intentionality concerns the temporal monitoring of imitation. In order to synchronize their activities, the imitatee waits and the imitator runs, which means that simultaneous imitation is a shared goal for the partners.

To assert that the setting is not a bias (= identical objects may enhance imitation artefactually) but rather a condition for long-lasting exchanges, we conducted other studies where we compared the same dyads of 2-year-olds meeting in two different settings: a setting with only one set of twenty different objects, and a setting with two identical sets of ten objects. Results showed that identical objects were a condition for immediate imitation and imitation a condition for long-lasting exchanges. This finding led us to consider identical objects as co-referents and imitative activities as a very simple way to share a topic and to understand, refuse or accept the other's focus of interest and goal (Nadel et al. 1999). This can be seen as another contribution of imitation to the building of a TOM. And as a matter of fact, identical objects did not appear to facilitate topic sharing after speech mastery. Forty-two-month-olds did not behave like 30 month-olds: they did not privilege the use of identical objects as co-referents but instead built symbolic referents supporting their goals and intentions. Imitation appears to be a transitory communicative system disappearing when speech is mastered, but working as an essential milestone toward the understanding of other's interests and intentions.

In short, imitation works like a communicative system through complex combinations of both imitating and detection of being imitated. The coordination of these two facets of imitation allows the children to take turn by alternating the roles of imitator and imitatee, and to share topics, via similar instrumental activities directed toward the same goal, using the identical object as a co-referent, a special case of joint attention (Nadel et al. 1999). Taking turn and sharing topics are decisive experiences to feel like the other one and to feel the other one like me. In particular, synchronically doing the same with similar objects and similar goals might well be a very important opportunity to establish equivalence in mental states between what I feel and think and what

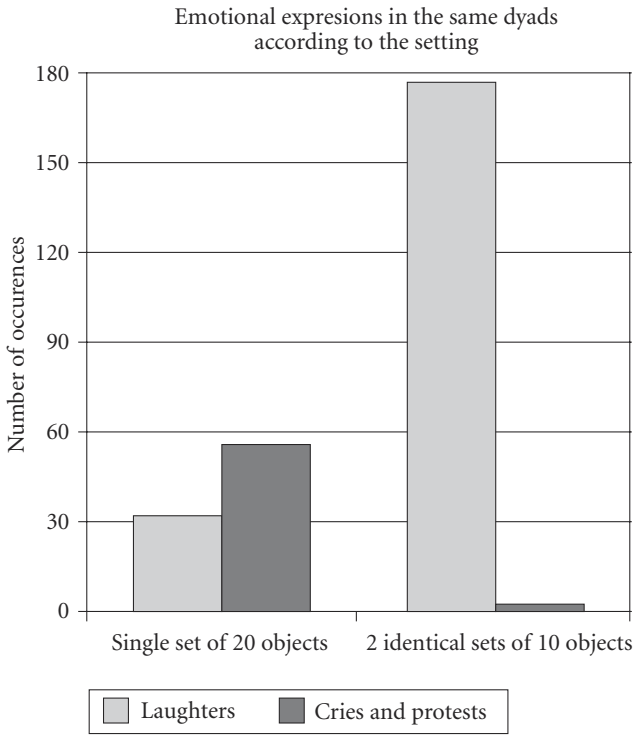

Figure 6. Emotions according to instrumental imitation availability

you feel and think. The number of positive emotional expressions displayed during these meetings is an important contribution to this suggestion.

Indeed, synchronous imitation and recognition of synchronous imitation provide a very special experience *on line*. This experience *on line* may allow infants to build a repertory of intentional concrete actions that can be attributed to oneself as well as to others: during an imitative episode, they can form not only an efference-copy of their own action, but also a copy of the other's action, since these actions are similar. Shared experiences of synchronous imitation may thus bring a major contribution to *off line*, flexible computation of primitive mental states.

Summing up

Imitation

1. informs about the understanding of persons vs. objects (neonatal imitation is selective)
2. concerns linkages between perception and action (action-perception coupling)
3. the specific behaviour does not matter, because the invariant in this situation is to "match"
4. implies mapping between notions of self and of other
5. suggests that others are "like me" (according to Metzoff & Moore's hypothesis of "interpersonal coupling")
6. forms sympathetic relating between me and other

Recognition of one's imitation by other

1. informs about the understanding of persons vs. objects
2. tests whether it is a random congruence or a systematic one
3. the specific behaviour does not matter, because the invariant in this situation is to "match"
4. suggests that seen acts are "like the one I do" (according to Metzoff & Moore's hypothesis of "interpersonal coupling")
5. suggests that self and other are equivalent agents, equivalent producers of intended acts
6. provides means to learn about who we are, and to form ontological expectancies

Notes

1. Reprinted from Nadel et al. (1999). *Developmental Science, 2*, 2, p. 167.

2. Reprinted from Nadel et al. (2000). *Autism*, 2.

3. To put it in Meltzoff and Moore's words (1999: 25).

4. *Ibid.*

References

DeCasper, A., & Fifer, W. (1980). Of human bonding: New-born prefer their mother's voices. *Science, 208,* 1174–1176.

Ellsworth, C., Muir, D., & Hains, S. (1993). Social competence and person-object differentiation. An analysis of the still-face effect. *Developmental Psychology, 29,* 1, 63–73.

Field, T. M. (1977). Effects of early separation, interactive deficits, and experimental manipulations on infant-mother face-to-face interaction. *Child Development*, *48*, 763–771.

Gergely, G. & Watson, J. (1999). Infant's sensitivity to imperfect contingency in social interaction. In P. Rochat (Ed.), *Early social cognition* (pp. 101–136). Hillsdale, NJ: Erlbaum.

Hains, S. & Muir, D. (1996). Effects of stimulus contingency in infant-adult interactions. *Infant Behavior and Development*, *19*, 49–61.

Jacobson, S. (1979). Matching behavior in the young infant. *Child Development, 50,* 425–430.

Legerstee, M., Pomerleau, A., Malcuit, G., & Feider, H. (1987). The development of infants' responses to people and a doll: implications for research in communication. *Infant Behavior and Development, 10,* 1, 81–95.

Maratos (1973). The origin and development of imitation in the first six months of the life. Paper presented at the British Psychological Society Annual Meeting, Liverpool.

Mehler, J., Jusczyk, P., Lamberz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young children. *Cognition*, *29,* 143–178.

Meltzoff, A. (1990). Foundations for developing a concept of self: The role of imitation in relating self to other and the value of social mirroring, social modelling, and self-practice in infancy. In D. Cicchetti & M. Beeghly (Eds.), *The self in Transition* (pp. 139–164). Chicago: University of Chicago Press.

Meltzoff, A. & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Flusberg, & D. Cohen (Eds.), *Understanding other minds* (pp. 335–366). Oxford: Oxford University Press.

Meltzoff, A. & Moore, M. (1983). Newborn infants imitate adult facial gestures. *Child Development, 54,* 702–709.

Meltzoff, A. & Moore, M. (1999). Persons and representation: Why infant imitation is important for theories of human development. In J. Nadel & G. Butterworth (Eds.), *Imitation in infancy* (pp. 9–35). Cambridge: Cambridge University Press.

Muir, D. & Hains, S. (1993). Infant sensitivity to perturbations in adult facial, vocal, tactile, and contingent stimulation during face-to-face interactions. In B. de Boysson-Bardies et al. (Eds.), *Developmental Neurocognition: speech and face processing in the first year of life* (pp. 171–185). Dordrecht: Kluwer.

Muir, D. W. & Nadel, J. (1998). Infant social perception. In A. Slater (Ed.), *Perceptual development: visual, auditory, and speech perception in infancy* (pp. 247–285). Hove, UK: Psychology Press Ltd. Publishers.

Murray, L. & Trevarthen, C. (1985). Emotional regulation of interaction between two-month-olds and their mother's. In T. M. Field & N. A. Fox (Eds.), *Social Perception in Infants* (pp. 177–197). Norwood, NJ: Ablex.

Nadel, J. (1986). *Imitation et communication entre jeunes enfants*. Paris: PUF.

Nadel, J. (2002). The functional use of imitation in preverbal infants and nonverbal children with autism. In A. Meltzoff & W. Prinz (Eds.), *The imitative mind: Development, Evolution and Brain Bases* (pp. 62–82). Cambridge: Cambridge University Press.

Nadel-Brulfert, J. & Baudonnière, P. M. (1982). The social function of reciprocal imitation in 2-year-old peers. *International Journal of Behavioral Development*, *5*, 95–109.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., & Reserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental Science*, *2*, 164–174.

Nadel, J., Guerini, C., Pezé, A., & Rivet, C. (1999). The evolving nature of imitation as a transitory means of communication. In J. Nadel & G. Butterworth (Eds.), *Imitation in Infancy* (pp. 209–234). Cambridge: Cambridge University Press.

Nadel, J. & Pezé, A. (1993). What makes immediate imitation communicative in toddlers and autistic children? In J. Nadel & L. Camaioni (Eds.), *New perspectives in early communicative development* (pp. 139–156). London, NY: Routledge.

Nadel, J., Croué, S., Mattlinger, M.-J., et al. (2000). Do children with autism have ontological expectancies concerning human behaviour?: A pilot study. *Autism, 2*, 2.

Reyes, L., Striano, T., & Rochat, P. (1998). Determinants of the Still-Face phenomenon by 2 to 6 month-old infants. Poster presented at the 11th Biennial International Conference on Infant Studies.

Schaal, B. (1988). Olfaction in infants and children. *Chemical Senses*, *13*, 145–190.

Snow, K. (1998). Social contingency and optimal social structures. *11th Biennial International Conference on Infant Studies.* Atlanta, Georgia, 2–5 April.

Soussignan, Schaal, Marlier, & Jiang (1997). Facial and autonomic responses to biological and artificial olfactory stimuli in human neonates: Re-examining early hedonic discrimination of odours. *Physiology and Behavior*, *62*, 745–758.

Stern, D. (1977). *Mère et enfant, les premières relations.* Bruxelles: Mardaga.

Trevarthen, C., Kokkinaki, T., & Fiamenghi, G. A. (1999). What infants' imitation communicate. In J. Nadel & G. Butterworth (Eds.), *Imitation in infancy* (pp. 127–185). Cambridge: Cambridge University Press.

Tronick, E., Als, H., Adamson, L., Wise, S., & Brazelton, T. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of American Academy of Child Psychiatry*, *17*, 1–13.

Uzgiris, I., Vase, M., & Benson, J. (1984). A longitudinal study of matching activity in mother-infant interaction. *Infant Behavior and Development, Special Issue, 7*, 371.

Zazzo, R. (1957). Le problème de l'imitation chez le nouveau-né. *Enfance, 10,* 135–142.

Reply to Jacqueline Nadel

Johannes Roessler
University of Warwick

1. Introduction

When reading about some of the recent experimental work on early imitation one may be forgiven the impression that imitation may matter a great deal more to developmental psychologists than to infants. It can look as if imitation is something which infants understand they are expected to do when the psychologists visit the creche, but which is of little importance in real life. One thing that struck me about the work presented in Jacqueline Nadel's chapter is just how thoroughly it refutes that impression. Partly, this is because much of her work is concerned with peer imitation and imitation with caregivers in naturalistic settings. But another reason, I think, is her emphasis on the question of what is the function of early imitation behaviours, and what kinds of motivation inform them. It seems to me that if we are interested in connections between early imitation and infants' understanding of other minds, then this is precisely the question we should be focussing on.

It is worth distinguishing at the outset two kinds of claims developmental psychologists have made about the relation between imitation and understanding other minds. First, there are claims to the effect that certain imitative behaviours manifest a certain kind of grasp of the idea of other minds. A somewhat provocative example is the suggestion that imitative behaviours shown by newborn babies are underpinned by a primitive 'theory of action', linking internal states like proprioceptive sensations to overt behaviour (Gopnik & Meltzoff 1997: 130). Second, there are claims to the effect that imitation constitutes a developmental mechanism for acquiring, or further developing, a conception of other minds. Here are two examples of this latter kind of claim. According to A. Meltzoff and A. Gopnik, mutual imitation provides children 'with a kind

of "primer in common sense psychology", a private tutorial in person-related versus thing-related interaction' (1993: 355). And according to P. Hobson, 'it is partly through imitating others' actions that infants come to experience and in due course understand what it means to assume another person's viewpoint' (1993a: 526).

A central theme in Nadel's work is the idea that imitation in infancy should be seen as a form of communication. Specifically, she suggests that around the middle of the second year imitation begins to serve a distinctively communicative function, a function that comes to a peak around 30 months and peters out towards the end of the fourth year. What I want to do in what follows is to explore how this suggestion bears on the issue of the relation between infants' imitation behaviours and their grasp of the idea of other minds.

One question here is what kind of psychological understanding may be manifested by imitation behaviours that are communicative in Nadel's sense. But there is also a more general issue regarding the relation between, on the one hand, the intuition that infants' imitation behaviour amounts to a form of communication and, on the other, the claim that it manifests a grasp of the idea of other minds. It is of course possible that the two ideas are quite unrelated (I think this view is implicit, for example, in Meltzoff's and Gopnik's account of the developmental role of imitation, to which I will return below). But it is certainly not absurd to think that the two thoughts might be closely connected or interdependent. Thus it might be argued that it is only insofar as (and in virtue of the fact that) imitation has a communicative character that we are entitled to regard it as manifesting a grasp of the idea of other minds. There are two different theoretical considerations that might motivate such a view. It might be held that communication provides a particularly significant source of evidence for someone's grasp of the idea of other minds. Or one might hold the view that being able to communicate with others is part of what it is to grasp the idea of other minds (or at least to grasp it in the way in which we normally do). There is no room here for a detailed general discussion of these claims. Rather I want to consider what I think is an interesting challenge for the view that communication is necessary for the idea of other minds: there seems to be a way of being aware of other minds that is available prior to the emergence of communication.

2. Nadel on imitation and communication

Suppose you put on your sunglasses, and as a result of watching you doing this I put on mine. And suppose further that my action is intentional, and intended as an imitation of yours. (It is not like the case of finding oneself yawning as a result of seeing someone else yawn.) The content of my intention may be more or less complex. The most basic case would be this: I intend to do as you do (regardless of what intentions, if any, inform your action). Let us simply call this intentional imitation. Again I may intend to do as you *intend* to do, or as you do intentionally. Call this psychological imitation. Finally, I may intend to do as you *intend for me to do*. That is, I may interpret your behaviour as a kind of message: 'this is what I want you to do'. Call this communicative imitation.[1]

I think one of Nadel's central claims may be put by saying that around the age of 18 months infants begin to engage in communicative imitation in this sense. (See her contribution to this volume; Nadel et al. 1999; and Nadel in press.) Support for this claim (and for this interpretation of her claim) comes from two main sources.

First, Nadel's findings suggest that around 18 months infants begin to play imitative games with their peers or caregivers involving what she calls reciprocal imitation: they respond to being imitated by going on to imitate the imitator (not by just repeating the same action, but by imitating an action that was not itself an act of imitation). So not only are they able to recognize being imitated, but they can shift the roles of imitator and model.

Nadel notes that this differs significantly from mere turn taking, where the participants simply alternate performing the same action, potentially indefinitely and with no separation of the roles of imitator and model (see Nadel et al. 1999). Even mere turn taking involves a form of coordination: the participants jointly regulate the order and the 'beat' of their respective contributions. (Trevarthen 1993) But reciprocal imitation requires, in addition, that the participants share a conception of who is imitator and who is model. If they cannot agree on this, their interaction will quickly grind to a halt. It is natural to think that such agreement is established by expressing and recognizing intentions, either to imitate an action or to perform an action to be imitated. Thus the participants' behaviour will be intended, and interpreted, as carrying a message, e.g. 'it's now my turn to be the model, this is what I want you to do'.

The second relevant finding concerns the point of imitation behaviours. Imitation is often regarded as a crucial mechanism for cultural learning (Tomasello et al. 1993). For example, imitating someone putting on sunglasses is an effective way of learning how to use sunglasses. But Nadel suggests that

towards the end of the second year imitation also begins to serve a purely social purpose, for example, as a way of expressing interest in someone, or simply as an enjoyable joint activity. Thus, 2-year-olds sometimes imitate clearly unconventional uses of objects (using sunglasses as earrings) or manifestly unsuccessful attempts to achieve a goal. So the point of imitative games is not exhausted by their role as an exercise in cultural learning. Perhaps the most interesting illustration of the distinction between imitation as cultural learning and imitation as communication is the case of deferred imitation. In Nadel's view, the point of deferred imitation is not exhausted by its obvious benefits in terms of being able to use objects appropriately. Rather, deferred imitation can also serve the purely communicative function of sharing memories (Nadel et al. 1999).

This account of 2-year-olds' imitation behaviours suggests that such behaviours are underpinned by quite sophisticated conceptual abilities. In particular, they require the ability to recognize intentions in others. Let me highlight three features of this ability.

a. The success of a game of reciprocal imitation depends on both participants pursuing different (though complementary) intentions. They must be able to recognize intentions that differ from their own current intentions.
b. Nevertheless, intentions can be shared. In fact, it is precisely in order to share an intention, say, of playing at imitation, that the participants need to adopt their different, complementary intentions of being imitator or imitatee. Jointly pursuing a project in this way involves exchanging reasons: by putting on my glasses, I mean to give you a reason to put on yours. We must both be, and take each other to be, responsive to such reasons.
c. Communicative imitation requires understanding, and acting upon, someone else's thoughts about oneself; for example, understanding that she wants me to imitate her action. So it seems to require what is sometimes called objective self-awareness, the ability to think of oneself as an object, or to have first-person thoughts.

To attribute these capacities to 2-year-olds is not, I think, to make a very controversial claim. There is a considerable amount of evidence to suggest that the ability to think of oneself as an object normally emerges around the middle of the second year (see especially Kagan 1981) – about the same time at which children are said to become able to reflect on the intentions informing (their own and others') actions (Perner 1991). In line with the general view mentioned at the end of the last section, it might be argued that children's emerging capacity for communication is partly constitutive of these conceptual abilities.

Thus the thought might be that 2-year-olds' notion of intentional action includes the idea that such actions are performed for reasons; and it might be argued that reasons, in the relevant sense, are essentially shareable and hence communicable. I won't discuss this line of thought in general terms. Rather, I would like to test it by considering an obvious objection. It is a well established finding that 1-year-old children, and even new born babies, engage in imitation. And many would argue that such early imitation behaviours already manifest some grasp of the idea of other minds. If this is right, it might seem to suggest that such grasp is available prior to the emergence of communication.

3. Newborn imitation

In a series of experiments conducted by A. Meltzoff and K. Moore, very young infants – some of them less than an hour old – were found to imitate simple facial gestures such as tongue protrusion or mouth opening (see Meltzoff & Moore 1995 for a review). It looks as if perceiving certain kinds of bodily movements immediately enables infants to execute similar movements. Meltzoff and Moore hypothesize that such early imitation behaviours exploit a kind of common coding, which they label a supramodal body scheme. Infants' perception of others' bodily movements and the planning of their own movements exploit the same mental code, so similarities between movements represented in these two ways are transparent.

This hypothesis raises interesting questions; for example, whether the claim is one about the content of infants' perceptual experiences and intentions, or one about sub-personal mechanisms underpinning their ability to match perceived and performed actions. But here I want to focus on Meltzoff's and Gopnik's contention that newborn imitation is 'nature's way of solving (...) the problem of other minds' (1997: 131).

We can distinguish two elements of this putative solution. (1) Newborn imitation reveals an innate capacity to link perceived behaviour with internal mental states like proprioceptive sensations and motor plans, thus providing a 'groundwork for the infant's theory of action' (1997: 133). (2) In perceiving others' movements in a way that makes them immediately imitable, infants experience others as being similar to themselves; their developing interpersonal relations are founded on the 'perceptual judgement "Here is something like me"' (1993: 336).

It seems to me that there are grave problems with both of these suggestions. Concerning (1), the critical question is whether newborn imitation re-

quires a *representation* of mental states such as kinaesthetic sensations or motor plans. On the face of it, it merely requires *forming* certain kinds of motor plans in response to perceived movements, and perhaps *experiencing* corresponding kinaesthetic sensations. Accordingly, the common coding hypothesis is naturally read as a claim about the representation of bodily movements in different modalities, rather than as a claim about the representation of mental states. True, in one sense infants 'link' another's movements and their own motor plans. But we should be wary of conflating that sense with the claim that they connect a perceived movement with *the other person's* kinaesthetic sensations and motor plans.

Concerning (2), there is an issue about self-awareness. The thrust of Gopnik's and Meltzoff's position is that newborn imitation undermines traditional conceptions of young infants as radically egocentric, since it reveals a recognition of basic similarities between self and other. But we need to draw a distinction between two ways of recognizing such similarities. It might be said that facial gestures provide infants with a kind of affordance, in Gibson's sense: the gestures are perceived as opportunities for action, as affording imitation, or, in short, as imitable. Now, which bodily movements you can imitate will of course depend on whether you are relevantly similar to the model, on whether you have the requisite limbs, motor skills and dimensions. But it seems plausible that infants may perceive a movement as imitable without reflecting on any of these similarities, and without being able to explain their own or someone else's imitation skills in terms of the relevant match. Their grasp of what it is to be imitable may be entirely practical – it may just be a matter of responding to a movement perceived as imitable by trying to imitate it. Note that such a notion of imitability does not require the ability to think of oneself as an object; it is not a notion that can be used in thinking about others' imitation skills, or to reflect about oneself as one imitator amongst others. In perceiving something as imitable one may exploit similarities between oneself and the other without any explicit recognition of them.

In other words, infants' notion of imitability may be causally indexical, in J. Campbell's sense (Campbell 1993). What makes a term causally indexical is the way in which the causal significance of the property in question is grasped. For example, perceiving an object as being within reach may just be a matter of perceiving the object as affording reaching for. Grasp of the causal significance of the fact that the object is within reach may be exhausted by grasping its implications for one's own actions. It need not involve the ability to give causal explanations of one's own, or others', actions citing the property. Indeed, one need not be able to think of objects as being within reach of someone else

at all. Most importantly, causally indexical notions do not require the ability to think of oneself as an object, or to have first-person thoughts. One may perceive information that is implicitly 'self-specific', such as that an object is 'within reach', without the ability to think of the object as being 'within my reach'. This latter kind of thought would involve the notion of something being within the reach of *x*, and that notion need not be available to someone with a purely practical grasp of what it is for an object to be within reach.

The advantage of this interpretation is that it does not require crediting newborn babies with objective self-awareness. But at the same time, and, on the face of it, connectedly, it suggests that early imitation behaviours are consistent with a certain kind of egocentrism. Infants may be aware of others as affording imitation, without thinking of imitation as an interaction between entities of the same type.

4. Imitating goal-directed actions and recognizing imitation

This interpretation of newborn imitation might prompt the following general hypothesis. Perhaps imitation behaviours in infancy generally fall into either of two categories. Imitation may be either a matter of exploiting affordances, without any explicit grasp of the similarities between the perceived and the performed actions, and indeed without explicit self-awareness at all. Or else it may involve genuine communication, with a grasp of communicative intentions, and awareness of oneself as one person among others.

Note that to say that imitation is a matter of exploiting affordances is not necessarily to deny that it involves a grasp of the other's psychological properties. Rather, the claim would just be that even when it does, grasp of such properties will merely consist in appreciating their significance *for one's own actions*, and will, in that sense, be egocentric. The following example of Nico Frijda's provides an illustration: 'understanding a given reaction as anger may consist of readying one's self to receive attack' (1969). There is some understanding of the causal role of anger here, of the fact that anger tends to trigger aggressive behaviour. But the causal role of someone else's anger is understood solely in terms of its significance for one's own actions.

So the hypothesis is that imitation behaviours are either properly communicative or else merely egocentric. We can put the hypothesis to the test by considering two abilities that are more sophisticated than newborn imitation, without, on the face of it, involving communication in the normal sense.

First, at least by 9 months, infants imitate novel object-directed actions, even after delays of several days. Thus, imitation begins to serve as a mechanism for cultural learning. As Meltzoff and Gopnik put it, 'children treat adults as a source of information about objects' (1993: 343). The question is what kind of psychological understanding, if any, deferred imitation reveals. Meltzoff and Gopnik take such imitative behaviours to be a case of what I earlier called psychological imitation, imitation aimed at doing as the other *intends* to do. They write: 'The 9-month-old infant seems to think that actions are directed toward objects in a goal-directed way, rather than simply conceiving of them as bodily movements per se.' (1997: 145). But if they are right about the point of such behaviours (that it consists in the transmission of information about objects), then it is tempting to think that 9-month-olds' grasp of the model's goal is, once again, purely practical. The model's behaviour may simply make it apparent to infants what to do with a particular object. While this involves some grasp of the goal informing the model's action, that grasp may just consist in letting the goal direct one's own actions, and, in that sense, still be egocentric. It is not clear that imitating object-directed actions reveals a detached understanding of others' goals as explanatory of their behaviour.

Second, there is evidence that by 14 months, infants are able to recognize being imitated. They seem to enjoy looking at an adult imitating them, and perform movements apparently designed to test whether they will be imitated. A natural suggestion would be that the children realize: 'she is trying to imitate my movements'. So the test behaviour may be intended to confirm this interpretation of the adult's behaviour. However, this account may be excessively reflective. Perhaps all that is going on is that the child perceives the adult's behaviour as 'controllable', i.e. as affording a peculiar kind of control, without any physical contact. It may be possible to perceive behaviour as controllable in this way without understanding the causal mechanisms underpinning the control. So the test behaviour may be another case of affordance-driven action.

The moral one might draw from this discussion is that prior to the development of communication, imitation behaviours manifest at best a purely practical grasp of other minds. I certainly think that this line of argument has some polemical value, as it challenges a conception of imitation (such as Gopnik's and Meltzoff's) as a mechanism which allows children to infer the internal states responsible for others' behaviour. Nevertheless, I strongly suspect that the stark alternative envisaged by our hypothesis – either full-blown communication or else mere egocentricity – distorts infants' intersubjective understanding. Let me end by sketching, if only in crudest outline, what I think may be a more plausible alternative. This alternative is certainly incompatible with

the letter of the view that the idea of other minds has its home in the context of communication; but I think not with its spirit.

5. Imitation as joint attention

M. Tomasello writes: 'Imitative learning has not been typically considered a relevant phenomenon when researchers talk of joint attention, but, obviously, it involves a following into adult behavior and attention in some sense' (1995:110). Joint attention, as the term is normally used by developmental psychologists, refers to a number of behaviours that emerge towards the end of the first year, for example social referencing or pointing behaviours. What they have in common is that the two participants perceptually attend to some significant object in their environment, and are in some sense aware that they share a common focus of attention. Now, if you imitate my putting on my sunglasses, you will probably focus your attention on your sunglasses, while my attention will be focussed on mine. Nevertheless, there is a sense in which we share a common focus of attention: the two of us are attending (as we say) to the same kind of activity, and we may be aware that we are. I want to suggest that thinking of early imitation behaviours as a form of joint attention may help to clarify the sense in which they manifest intersubjectivity.

It is helpful, first of all, to reflect on a distinction psychologists have drawn between two kinds of pointing behaviours. Sometimes infants point to a desired object in order to enlist an adult's help to obtain the object. This is referred to as proto-imperative pointing. In contrast, proto-declarative pointing is a matter of directing an adult's attention to some interesting object, apparently just in order to (in some sense) share one's appreciation of the object. There is a powerful intuition that while in proto-imperative pointing other people may conceivably just be treated as complex instruments (with facts about others' focus of attention understood in terms of their consequences for one's own instrumental actions), proto-declarative pointing reveals an awareness of the other as a subject of a point of view. Put crudely, such an awareness seems to be integral to the *point* of proto-declaratives: the project of sharing attention would be impossible to pursue without it.

A similar distinction may be applied to imitation. Imitating object-directed actions may be aimed at gathering desired information as to how to use a particular object. But it may also serve the purpose of sharing attention, of jointly pursuing an interesting activity. This distinction is, of course, reminiscent of Nadel's distinction between imitation as cultural learning and imitation

as communication. However, there is an important difference. Joint attention is a less sophisticated achievement than communication proper. Unlike communication, it may not require the ability to interpret others' behaviour in terms of their intentions, nor the ability to think of oneself as an object. As Werner and Kaplan put it in a suggestive formulation, early forms of social interaction 'have the character of "sharing" experiences *with* the Other rather than "communicating" messages *to* the Other' (1963: 42).

It is of course none too clear what is involved in such sharing of experiences. Once again, there is a temptation to think that it must be a matter either of explicitly recognizing the similarities between oneself and the other, or of a merely practical, hence egocentric, grasp of the other as affording desirable activities. But I think it must at least be a move in the right direction to insist that joint attention may involve a practical, yet non-egocentric grasp of other minds. It is practical in the sense that the causal significance of others' mental properties is grasped in terms of the way they affect one's own projects, rather than in terms of the way they may help to make the others' behaviour intelligible. Yet it is non-egocentric, given that the project is precisely that of sharing experiences. The challenge here is to specify the content of that project, as infants conceive it, in terms that do not presuppose self-awareness or a detached grasp of others' intentions.

In a series of papers, P. Hobson has argued that children's developing grasp of the idea of other minds is based on their ability to share affective attitudes with others. He makes two specific claims, which I think can be usefully brought to bear on the case of early imitation. One is that infants become first familiar with others' psychological states by becoming perceptually aware of others' affective attitudes. Most importantly, in the current context, according to Hobson, around the age of 9 months infants become able to perceive that others have object-directed emotions. In other words, the intentionality of the mental is first encountered in the intentionality of affective attitudes. The other claim concerns the nature of infants' awareness of such states: their responsiveness to others' affective attitudes is partly an *affective* responsiveness. Infants first become aware of intentionality in the context of sharing attitudes, where they respond to a perceived attitude to an object by adopting it (Hobson 1993a, 1993b).

Suppose we grant, as I think we should, that affective attitudes can be directly apprehended in people's facial expressions and expressive behaviour. We might then say that the goal informing object-directed imitation may be that of sharing the other's manifest enjoyment of a particular action; just as the goal informing proto-declarative pointing may be that of sharing pleasure or alarm

about certain features of an object. So the project of sharing experiences does not require a sophisticated conception of experiences as explanatory of someone's behaviour. Furthermore, it does not even require the ability to reflect on one's own experiences. Awareness of a shared focus of (affectively coloured) attention may be, as it were, immersed rather than reflective. It may consist, not in recognizing that she is paying attention to the object to which I am also attending, but just in being aware that her attention is focussed on 'this object' or on 'this activity' (and on certain affectively significant features of them), where the demonstrative identification involves focussing one's own attention on the object or activity.

It might be said that appeal to affective attitudes in connection with imitation does not really advance matters. For it is hard to see how one could be aware that someone is attending to an action, or attentively doing something, without having at least some idea of the point of her action – of the intention informing her action and controlling the focus of her perceptual attention. In the end, if imitation involves a sharing of mental states, it will have to be a sharing of intentions. I think that in a sense, this is true. But the question is what is involved in recognizing another's intention. Following Hobson's lead, it might be argued that infants grasp the intention informing someone's action precisely in being perceptually aware of the agent's affective attitude, e.g. by recognizing which action it is she is enjoying. Recognizing such enjoyment does not require the ability to reflect on the ends someone might be pursuing in using certain means, and to explain actions in terms of the agent's intentions.

Interestingly, on the present account, one developmental version of the problem of other minds simply vanishes – the problem of how infants work out that others too have the kinds of conscious states with which they are familiar from their own case. (This seems to be the problem Meltzoff and Gopnik have in mind when they aver that imitation is 'nature's way of solving the problem of other minds' (1996: 131).) What infants acquire first (towards end of the first year), on the present account, is a notion of *others* as subjects of a point of view. The insight they must then achieve, on the basis of this, is that they themselves are entities of the same type as others, which in turn enables them to think of others as subjects with their own first-person perspective.[2] To the extent that imitation is one of the mechanisms facilitating this development, it would seem to be a mechanism for the development of self-awareness as much as for the development of the idea of other minds.[3]

Notes

1. For imitation to be communicative it must, in addition, meet the (intuitively compelling, but hard-to-analyze) condition of mutual openness. Very roughly, the model must have no intention to deceive the imitator about the intentions informing her action, and the imitator must not credit the model with such an intention. For example, if you take it that I think that you are unaware of my observing your action, then while you may still intend your action to be seen as an invitation to imitate it, it would not be informed by a communicative intention.

2. This conception of the developmental trajectory has been proposed by P. Hobson: see e.g. his 1993a and 1993b.

3. I would like to thank Joëlle Proust for helpful comments on an ealier draft.

References

Campbell, J. (1993). The role of physical objects in spatial thinking. In N. Eilan et al. (Eds.), *Spatial Representation*. Oxford: Blackwell.

Frijda, N. (1969). Recognition of Emotion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, Vol. 4. New York: Academic Press.

Gopnik, A. & Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Hobson, P. (1993a). On acquiring the concept of "persons". *Behavioral and Brain Sciences, 16*, 525–526.

Hobson, P. (1993b). The emotional origins of social understanding. *Philosophical Psychology, 6*, 227–249.

Kagan, J. (1981). *The Second Year. The Emergence of Self-Awareness*. Cambridge, MA: Harvard University Press.

Meltzoff, A. & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding Other Minds*. Oxford: Oxford University Press.

Meltzoff, A. & Moore, K. (1995). Infants' understanding of people and things: From body imitation to folk psychology. In J. Bermudez, A. Marcel, & N. Eilan (Eds.), *The Body and the Self*. Cambridge, MA: MIT Press.

Nadel, J., Guerini, C., & Rivet, C. (1999). The evolving nature of imitation as a transitory means of communication. In J. Nadel & G. Butterworth (Eds.), *Imitation in Infancy*. Cambridge: Cambridge University Press.

Nadel, J. (2002). The functional use of imitation in preverbal infants and nonverbal children with autism. In A. Meltzoff & W. Prinz (Eds.), *The Imitative Mind: Development, Evolution, and Brain Bases*. Cambridge: Cambridge University Press.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Tomasello, M. (1995). Joint Attention as Social Cognition. In C. Moore & P. Dunham (Eds.), *Joint Attention*. Hove: Erlbaum.

Tomasello, M., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences, 16*, 495–511.

Trevarthen, C. (1993). The self born in intersubjectivity: The psychology of an infant communicating. In U. Neisser (Ed.), *The Perceived Self*.

Werner, H. & Kaplan, B. (1963). *Symbol Formation*. New York: Wiley.

Chapter 6

Varieties of simulation

François Recanati
Institut Jean-Nicod, CNRS, Paris

1. Assumption as the simulation of belief

The belief faculty can be seen as a mental 'module' consisting of two things: the mental encyclopedia or belief box (a set of representations), and an inferential device (Sperber 1985: 54). The function of the belief box is to store true representations of the world, which can then be used to guide action. The function of the inferential device is to exploit the representations in the belief box, that is, to derive consequences from them – to use them as premisses in theoretical and practical reasoning.

In order to ensure that representations stored in the belief box are true, there are input conditions, i.e. conditions which a representation must satisfy before it can enter the belief box. A representation can be tokened into the belief box only if it is validated, that is, only if the tokening results from a reliable cognitive mechanism – a causal mechanism which normally delivers true representations as outputs. Among the mechanisms in question we find perception and communication. Inference is also a validating mechanism: by feeding a representation from the belief box to the inferential device, we obtain further representations which are (inferentially) validated and can therefore go into the belief box.

Through exploitation, the belief box can be enlarged, because representations which are inferred from representations already in the belief box are themselves 'validated'. Exploitation is also what enables us to check whether a candidate for the belief box (i.e. a representation delivered by a reliable mechanism) is OK and yields no contradiction with representations already in stock.[1]

Acceptance can be defined as *the disposition to exploit a representation*, i.e. to feed it to the inferential device. Representations in the belief box are *eo ipso*

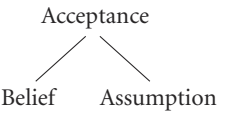

Figure 1.

accepted: When a representation is in the belief box, the cognitive agent is automatically disposed to exploit it. It is in virtue of this disposition that representations in the belief box freely interact with each other. But there are cases of acceptance which are not cases of belief. The cognitive agent may be disposed to exploit a representation which does *not* satisfy the input constraints, for example a representation which is not validated. The representation is accepted but it does not go into the belief box. Such a representation is merely *assumed*.

As I have defined it, acceptance is the most general notion (Figure 1), but the notion of belief is primary. We need the notion of acceptance, broader than that of belief, only because *within the belief module* there is a sub-system (the inferential device) which can be used to do something other than fulfill its function within the module (exploit representations in the belief box). The sub-system is individuated in part by its function within the belief system, hence the notion of acceptance, *qua* disposition to run the sub-system in question, presupposes the notion of a complete belief system.[2]

In assumption, the inferential mechanism is fed an input distinct from its normal input. This makes assumption a case of *simulation*, in the technical sense. There is simulation whenever a mechanism M which has a function within a system S, and exists only because of that function, is used 'off-line' (Nichols et al. 1996; Currie 1995), 'outside' the system. Such a use is parasitic on the normal use (as the mechanism exists only in virtue of its normal use). The belief faculty is such a system S with two components whose functions are determined within the system: the function of the belief box is to store the sort of representation (viz. true or at least validated representations) which it is the function of the inferential device to exploit. In this framework, assumption can be characterized as involving a divergence between the natural function of the inferential component (to exploit validated representations) and the way it is actually used (to exploit representations which are *not* validated).

Assumption is simulative not only in the technical sense but also in a more phenomenological sense. When someone assumes that *p*, he or she behaves *as if* the representation that *p* was validated, while in fact it is not. A classical example of assumption is the lawyer who believes her client guilty but assumes his innocence while defending his cause. For the lawyer to assume this is for her

to *pretend to believe* that her client is innocent. Another well-known example is hypothetical reasoning. In hypothetical reasoning, we pretend that something is the case, and we reason from that assumption. The representation assumed is fed to the inferential mechanism, and consequences are derived. This is like running a simulation.

Reasoning from assumptions is at the heart of the natural deduction approach to logic. In natural deduction, the act of 'assuming' is the act of *simulating assertion*. Propositions assumed are propositions which are fictively asserted. From those propositions consequences are derived, which inherit the fictitious character of the assumptions from which they are derived: they are not actually asserted, because they are derived from propositions which themselves are not asserted, but assumed (simulated). In other words the consequences are derived only 'under assumption', that is, simulatively. Yet the simulation establishes that such and such consequences 'follow from' such and such assumptions; and this fact concerning the implication relations between propositions can be genuinely asserted.

It is sometimes claimed that the form of simulation involved in reasoning from assumptions is a sub-personal procedure much like the building of a mental model in carrying out an inference; the point being that a sub-personal procedure of that sort has nothing to do with the personal process of 'pretending' (Sperber, in conversation). But there is an obvious continuum between that form of simulation and other forms which clearly engage the subject. Given this continuity, the burden of proof lies on the side of those who want to posit a sharp division between the two sorts of case.

Consider children's games of make believe. Two children pretend, say, that a certain stick is a sword. If this is not pretense in the personal sense, what is? What the children do is create a common disposition to exploit a representation which is not validated. They pretend in the sense that they do as if the representation was validated, and act accordingly. They assume that the stick is a sword.

From children's games of make believe to fiction there is but a short step, as Walton and others have shown (Walton 1990; Currie 1990). The consumers of a fiction assume the 'truths' of the fiction, i.e. exploit them as if they were validated. More precisely, the participants in the fictional game (the producer and the consumer) pretend that the representations of the fiction are *communicationally validated*: they pretend that the author is telling known facts. As Lewis says, 'the storytellers pretend to pass on historical information to their audience; the audience pretends to learn from their words, and to respond accordingly' (Lewis 1978: 276). That the pretense in question takes place at the

personal level is shown, inter alia, by the well-known fact that fiction involves the affects of the audience.

Among the forms of simulative assumption which involve affects and the emotions we find imaginative or counterfactual phantasies, to which fiction is obviously related. Kahneman and Tversky have studied those phantasies which they construe as 'the running of a simulation model' (1982: 201).[3] They claim that there are 'rules of mental simulation', which must be experimentally investigated and which govern counterfactual thinking. The mental simulation they talk about is responsible for the emotions which we associate with imagined situations: 'counterfactual emotions', as Kahneman and Tversky call them (1982: 206).

Another important family of cases, some of which are also affect-involving, concerns the simulation of the psychological states of other persons. There are two sub-cases. In *empathy*, knowing the situation of another person, we simulate and predict that person's emotional attitudes and reactions. In the other, more intellectual type of case we simulate and predict the *propositional* attitudes of a person *x*, by assuming the representations which (we know) *x* accepts and feeding them to the inferential mechanism. That is the process of *simulative belief ascription*, which has been the focus of much research in recent years. It provides a good example of the first form of simulation I want to talk about in this paper: simulation as a method for exploiting meta-representations.

2. Simulative reasoning as a method for exploiting meta-representations

Consider propositions (1), (2) and (3):

(1) Peter believes that John's phone number is 325–248.

(2) Peter believes that John has the same phone number as Mary.

(3) Peter believes that Mary's phone number is 325–248.

From (1) and (2) we can infer (3). This can be established experimentally. But how does the inference proceed? There are two options:

First option: We appeal to a bunch of rules of inference like:

[A] From Bel (x, p) and Bel (x, q) infer Bel (x, p & q)

[B] From Bel (x, p) and p → q infer Bel (x, q)

Such rules, and the general principle which underlies them, may be construed as defeasible heuristics employed in belief ascription, that is, as components of the "theory of mind". Using them, we can infer (3) from (1) and (2).

Second option: We run a simulation of Peter's beliefs. Each time a belief is ascribed to Peter, we *assume* that belief. Thus we have:

(1)	Peter believes that	John's number = 325–248	(premiss)
(2)		John's number = 325–248	(assumption)
(3)	Peter believes that	John's number = Mary's number	(premiss)
(4)		John's number = Mary's number	(assumption)

We can then reason 'within' Peter's belief space (Fauconnier 1985). From the two assumptions (2) and (4) we infer:

(5) Mary's number = 325–248

This is derived only under the assumptions (2) and (4). (5), therefore, cannot be asserted, but we can discharge the assumptions (2) and (4) and return to the 'parent space', in Fauconnier's terminology:

(6) Peter believes that Mary's number = 325–248

This simulative solution to the phone number problem has been discussed by various authors who have argued that it corresponds to the way we actually proceed. (Creary call it 'simulative reasoning', Dinsmore 'parochial processing'.) It involves two complementary processes: projection and retrojection. In our example, projection (steps 2 and 4) is the process of assuming, or more generally accepting, the propositions which Peter is said to believe; retrojection (step 6) is the process of ascribing back to Peter a belief in the propositions which follow from the assumed propositions. Simulative reasoning consists in projection followed by deduction followed by retrojection:

Inference X

(1)	Peter believes that	John's number = 325–248	(premiss)
(2)		John's number = 325–248	(1, projection)
(3)	Peter believes that	J.'s number = M.'s number	(premiss)
(4)		J.'s number = M.'s number	(3, projection)
(5)		Mary's number = 325–248	(2, 4, deduction)
(6)	Peter believes that	Mary's number = 325–248	(5, retrojection)

On this view we do not appeal to special rules of inference like [A] and [B], but we use a more general mechanism of *cognitive projection.* This mecha-

nism is not at work only in psychological explanation (belief ascription); it is used also in counterfactual phantasies, in fiction, and in a number of other cognitive domains.

An argument against the simulative approach runs as follows. If, when a meta-representation is asserted, we were free to assume the object-representation, to exploit it, and to "retroject" the consequences into the meta-representation, the following inference would be licensed:

Inference Y

(1)	Superman = Clark	Kent	(premiss)
(2)	Lois believes that	Superman is beautiful	(premiss)
(3)		Superman is beautiful.	(2, projection)
(4)		Clark Kent is beautiful	(1, 3, deduction)
(5)	Lois believes that	Clark Kent is beautiful	(4, retrojection)

But inference Y is incorrect: (1) and (2) are true, but (5) is false.

That is not a real counter-example to the simulation approach. There is a significant difference between inference X and inference Y. In X, (5) is deduced from two assumptions, while in Y, (4) is deduced from an assumption, viz. (3), together with a premiss which is *not* an assumption, viz. (1). (1) and (3) do not appear in the same 'column' in Y, and this is what makes the inference unacceptable. It is unacceptable because, in reasoning within Lois's belief space, we have used a premiss *external* to that space.

In simulative reasoning the temporary assumptions we make determine an 'assumptive space' within which we can reason: consequences derived within that space can be retrojected. But in so reasoning within the assumptive space we are not allowed to use any old premiss which holds outside the space. The only premisses which we are allowed to use in reasoning from assumptions are *premisses which presumably hold in the assumptive space at issue*. In particular, we cannot use a premiss incompatible with something in Lois's belief space, if we are to reason from an assumption in that space. But this is precisely what happens in inference Y. A premiss from outside Lois's belief space, namely the identity in (1), is used to derive consequence (4). Since this identity is incompatible with what Lois believes, it cannot be used to derive a retrojectible consequence. Hence inference Y is indeed incorrect. In contrast, the derivation of (5) in inference X does not appeal to premisses external to the assumptive space: the only premisses appealed to in deriving (5) are (2) and (4), which are themselves assumptions (and, as such, are written on an indented line).

That is not to say that we are absolutely prohibited from using external premisses in reasoning from assumptions. Let us define an external premiss as

one which results neither from projection nor from deduction from already available assumptions. Under certain conditions, we can (and often do) *copy* an external premiss into the assumptive space. Inference Y can be modified accordingly, through the introduction of an extra-step:

Inference Y*

(1)	Superman = Clark	Kent	(premiss)
(2)	Lois believes that	Superman is beautiful	(premiss)
(3)		Superman is beautiful	(2, projection)
(1′)		Superman = Clark Kent	(1, copy)
(4)		Clark Kent is beautiful	(3, 1′, deduction)
(5)	Lois believes that	Clark Kent is beautiful	(4, retrojection)

The simulative inference is now formally correct, but it is still unacceptable, because Copy is a legitimate move only if the proposition we copy into the assumptive space is compatible with propositions already in that space. (Under those conditions, Copy is the process which Fauconnier calls 'Optimization'.) But we have seen that (1) is incompatible with Lois's beliefs. The inference, therefore, is blocked at step (1′).

3. The simulative nature of meta-representations

As the phone number example shows, simulative assumption is a useful method for *exploiting* meta-representations. If we know what someone believes and/or desires, we can conduct inferences by assuming those desires and beliefs and seeing what results: which emotions, which intentions, which beliefs we would entertain if we were in their position, with their beliefs and desires. Insofar as we are similar to the agents whose thought, behaviour or emotion we are trying predict, we can rely on the simulation procedure.

But there is a much closer link between meta-representation and simulation. Not only can meta-representations be simulatively exploited by assuming the object-representation; meta-representations, I claim, are *intrinsically* simulative. It's not a matter of what we do with meta-representations, but a matter of what they are.

Meta-representations are not merely representations which are about representations – for example beliefs about other person's beliefs. 'John believes strange things' is a belief about John's beliefs, but it is not a meta-representation in the strong sense in which I am interested here. (It's a generalization over meta-representations, rather than a genuine meta-representation.)

In the strong sense, a meta-representation contains the primary representation, as the sentence 'John believes that grass is green' contains the sentence 'Grass is green'. In other words, a meta-representation is not merely a representation that refers to, or is about, a representation: it is a representation that refers to a representation *and* displays (part of) its content.

When I say that the meta-representation 'displays' the content of the object-representation (or an aspect of it), I mean that a sub-part of the meta-representation, which I call the radical, *actually expresses that content*. Consider the following example:

> Tom: – The Moon is made of green cheese
> Bob: – According to Tom, the Moon is made of green cheese

Tom's utterance represents the Moon as being made of green cheese. Bob's utterance, which is a meta-representation, has a different subject-matter: it is about Tom's utterance (or the belief it expresses), as the prefix 'According to Tom' makes clear. Yet Bob's meta-representational utterance displays the content of Tom's utterance and therefore replicates it to some extent. *Both Tom's utterance and Bob's include a representation of the Moon as being made of green cheese.* That representation is offered by Tom as a representation of the way the world actually is; and by Bob as a representation of the way the world is according to Tom.

To sum up, a meta-representation (in the strong sense) consists of two parts: one, the radical, is the object-representation itself, or at least it replicates it and expresses its content; the other is a tag (e.g. 'according to Tom') which situates the object-representation thus replicated. On this view, meta-representations have an *iconic* dimension (they resemble what they are about), and they can only be *transparent*. As undergraduates we learned that if a meta-representation *m* represents an object-representation *r* and *r* is about *x*, then *m* is about *r* and *not* about *x*. Meta-representations are supposed to be opaque. If what I have said is right, however, just the opposite is true: Whenever a meta-representation *m* displays some aspect *a* of the content of the object-representation *r*, if *r*, considered under aspect *a*, is about *x*, then *m* also is bound to be about *x*.

A genuine meta-representation dS (where d is the tag and S the radical) satisfies the following schema:

> *Schema (I):*
> One cannot entertain the proposition that dS without entertaining the proposition that S.

For example:

> One cannot entertain the proposition that John believes that grass is green without entertaining the proposition that grass is green.

Using Schema (I) as a criterion, we see, once again, that not all 'representations of representations' are genuine meta-representations. There is a difference between e.g. (1) and (2):

(1) Tom stated Leibniz's Law

(2) Tom stated that identical objects have all their properties in common

Both statements are about a statement, namely Tom's statement of Leibniz's Law. However, the content of Tom's statement (viz. Leibniz's Law) is displayed only in (2).[4] It follows that only (2) is a genuine meta-representation, that is, a meta-representation which displays the content of the object-representation and thereby replicates it. (1) is not iconic – it does not satisfy Schema (I): One *can* entertain the proposition that Tom stated Leibniz's Law without entertaining the proposition that identical objects have all their properties in common (= Leibniz's Law); but one *cannot* entertain the proposition that Tom stated that identical objects have all their properties in common without entertaining the proposition that identical objects have all their properties in common. Among genuine meta-representations, i.e. representations of representations which satisfy (I), we find the following:

> According to John, grass is green
> John believes that grass is green
> John says that grass is green
> In the picture, grass is green
> In the film, grass is green
> etc.

In virtue of the property I have just mentioned, one cannot entertain a meta-representation, for example a representation of John's belief, without also entertaining the putative content of John's belief. This entails that "thinking about others' thoughts requires us... to think about the states of affairs which are the subject-matter of those thoughts" (Heal 1998: 484). This is what I mean when I say that meta-representations are intrinsically simulative. The capacity of 'putting oneself into someone else's shoes', which is crucially involved in simulative reasoning, is already necessary in order merely to *entertain* a genuine meta-representation. Consequence: we cannot ascribe contents to a radically

alien creature. We can only ascribe contents to a creature which is sufficiently like us and whose conceptual resources are the same as ours.

4. Complex representations, 'entertaining', and simulation

The structure dS of meta-representations is not specific to them. There are lots of representations which have the same structure and consist of a primary representation and a tag. What I said of meta-representations is true of them too – or at least of some of them. For example, 'In Chicago, it is raining' is a complex representation which one cannot entertain without entertaining the primary representation 'it's raining'. Normally, the primary representation 'it's raining' gets tokened when the subject perceives rain around him or her. That is the function of that representation. In 'In Chicago, it is raining', the representation of rain serves a different purpose. It is used as a building block for constructing the complex representation. That requires decoupling the representation from the subject's experiential situation. That decoupling, arguably, is also a form of simulation.

According to the psychologist and language-theorist Karl Bühler, simulation is involved not only when we think or talk about imaginary situations, but more generally, whenever we think or talk about a situation that is not given to our senses. Following Bühler we can distinguish two basic representational modes, or rather two poles. In the *egocentric* mode what we say or think directly concerns the situation in which we presently are: the hic-et-nunc situation, as I call it. That mode is dominant in what we may call the 'egocentric stage' in the child's linguistic development, when only 'monoremes' or one-word utterances are issued (Sechehaye 1926). Such utterances are holistically predicated of the child's egocentric situation. Not being liable to vary, that situation is not and does not have to be articulated. As Perner says, "the perceiver has no option of representing anything but current reality" (Perner 1991: 67).

The other mode is more complex and involves what Bühler calls 'a liberating step' from the hic-et-nunc situation: the subject freely considers, and characterizes, situations which are not given to his or her senses – for example situations she remembers or imagines. According to Bühler what makes the liberating step possible is mental simulation: the absent situation is *presentified* through an act of the imagination.[5] (Even memory involves such an act.) As a result, the subject both experiences the egocentric situation and, at the same time, simulatively experiences the absent situation, which comes before 'the mind's eye'. Being simulatively presentified, the imagined situation

and what occurs in it can be deictically referred to much like the hic-et-nunc situation and what occurs in it can be deictically referred to. Bühler calls this 'imagination-oriented deixis' as opposed to (normal) perceptual deixis.

The liberating power of the simulation comes from the fact that the imagined situation can replace the egocentrically given situation as that which the current thought or discourse concerns. (In Perner's terminology, the thinker now uses 'multiple models'.)

> [If] a proper name such as 'Heidelberg' or 'Lake Constance'... is dropped as the subject of a sentence in the presence of a hearer who has been there, then an imagination-oriented deixis is at work and the release of the sense of the sentence from the support for understanding provided by the concrete speech situation has been more or less adequately prepared; if one is displaced in imagination to a thing, *one can forget from where one was displaced...*
>
> If I hear 'it is raining' without any prelude, I take this to be a diagnosis of the weather in the speech situation; it is raining where the speaker is at the moment... The addition of an exposition 'at Lake Constance' brings a releasing step: 'it is raining at Lake Constance'; this extended dictum can be spoken anywhere, its sense is largely liberated from... the speech situation.
>
> (Bühler 1934: 426)

I think Bühler is basically right. The function of representations is to represent the world around us. When a representation *r* is tokened, it depicts the world as being thus and so, in virtue of the general function of representations: to get tokened only if the environment is as the representation represents it as being. Now take the complex representation: 'Three miles from here, there is a blue castle with a high tower'. The complex representation purports to describe the world around the subject. The representation is true iff, three miles from where the subject is, there is a blue castle with a high tower. But the primary representation 'there is a blue castle with a high tower' does not purport to describe the world around the subject. There is no blue castle where the subject is. The primary representation purports to describe a situation *distinct from* the subject's egocentric situation. It is decoupled from the latter. Since the primary function of representations is to correctly describe the subject's egocentric situation, this tokening fulfils a deviant function. Instead of being used directly to provide information concerning what is going on, the token is used indirectly as a building block in the construction of a more complex representation which will provide information concerning what is going on.

The picture that emerges is this. In the basic egocentric mode, all tokenings concern the subject's experiential situation. They result from perceptual processes which are anchored in the subject's experiential situation (what's per-

ceived is what's going on here and now). Since tokenings result from such processes, the tokens indicate what the subject's egocentric situation is like. That is the function of mental representations in the egocentric stage. (As Dretske pointed out, that indicating function is what endows mental symbols with their representational content. A symbol represents what it is its function to indicate.) Now imagination enables us to go beyond the egocentric stage. Imagination is the faculty of simulating perception. Through imagination, tokenings are produced which do not result from actual perceptions. Such tokenings can still be fed to the inferential device: that is what happens in assumption. Simulative tokenings can also be used to represent a situation not given to the senses. One simulatively entertains a representation decoupled from the egocentric situation, while tagging it as a representation of what takes place three miles from here, or of what took place two days ago.

I see meta-representation as a special case of that mechanism. In meta-representation, the situation the simulative representation concerns is not a portion of the actual world distinct from the subject's own egocentric situation, but an *imaginary* situation, that is, a situation which belongs to a 'possible world' distinct from the actual world: a 'fictional world', a 'belief-world', etc. That difference between 'In the kitchen, S' and 'In the film, S' should not hide the striking commonalities. Note that simulative reasoning can be used not only to exploit meta-representations, but also to exploit any complex representations with the same general characteristics. If I know that in Chicago it is raining, I can infer that in Chicago the pavements are wet simply by appealing to the conditional 'if it is raining, the pavements are wet' (where no place is mentioned). By projection I go from 'In Chicago it is raining' to 'it is raining'. Using the above conditional I infer 'the pavements are wet'. By retrojection this yields: 'In Chicago, the pavements are wet'. See McCarthy for more on this type of inference.

5. Opacity

Let us go back to meta-representations. The view I have expounded entails that the ascriber shares, indeed must share, the ascribee's conceptual resources. Thus I can't ascribe content to someone whose concepts are too different from mine. That is why, as Geach once put it, *oratio obliqua* cannot turn nonsense into sense. If Heidegger's utterance, 'nothing noths', is nonsense, the report 'Heidegger believes that nothing noths' is still nonsense. (This I call the Inheritance Principle.)

But there are counterexamples.

> My son believes that I am a 'philtosopher' (where 'philtosopher' is a non-word and corresponds to a concept very different from the speaker's concept of a philosopher).
> My son believes that Santa Claus will come tonight.

Those examples seem to show that the ascriber need not share the ascribee's conceptual resources. There may be all sorts of conceptual divergence, including ontological divergences, between ascriber and ascribee. These divergences are generally accounted for in terms of opacity. Like quotation marks, operators like 'John believes that' are said to constitute "an opaque interface between two ontologies, two worlds: that of the attitudinist, however benighted, and that of our responsible ascriber" (Quine 1995: 356).

But that is precisely the view I opposed. I argued that meta-representations are fundamentally *transparent*: whichever state of affairs the object-representation represents, the meta-representation also represents. The ascriber and the ascribee are in the same boat since they both use the same representation. Bob, who says 'The moon is made of green cheese', and Tom, who reports 'According to Bob, the moon is made of green cheese' both offer a representation of the moon as being made of green cheese. For that to be possible, they must share the conceptual resources necessary to construct and entertain the representation in question. How, in this framework, can we deal with the well-known fact that meta-representational contexts are opaque in the technical sense? How can we deal with the examples in which there are manifest conceptual or ontological divergences between ascriber and ascribee?

According to Quine, a linguistic context is opaque if substitution of identicals is not truth-preserving in that context. Now there are two sorts of 'substitution of identicals', hence two sorts of substitution failure. One sort involves the interchange of expressions with the same *extension*. To account for that sort of substitution failure, it is sufficient to invoke the phenomenon of *circumstantial shift*.

The extension of an expression depends upon two factors: its content, and the situation or 'circumstance' in which we evaluate the expression. The extension of an expression always is its extension in or at a circumstance. Thus a description like 'the President' will denote Chirac or Clinton, depending on whether we evaluate it with respect to the French situation or the situation in the US. Now when we say that two expressions (e.g. 'Tegucigalpa' and 'the Capital of Honduras') 'have the same extension,' what we mean is that they have the same extension *in the current circumstance*, that is, at the present time and

in the actual world. But the tag in a meta-representation dS indicates that the internal sentence S must be evaluated with respect to a circumstance c' distinct from the current circumstance c in which the meta-representation itself is being evaluated. Hence it is only normal that we can't always substitute an expression for another one in the internal sentence even though they have the same extension *in the current circumstance*; for the current circumstance is irrelevant when we evaluate the internal sentence and the expressions it contains. What is relevant to evaluating the internal sentence is not the current circumstance c, but the shifted circumstance c'.

Let us now turn to the other, more problematic sort of substitution failure. If two expressions have the same content, they should be interchangeable *salva veritate*, even if the sentence is headed by a circumstance-shifting tag: for the content of an expression is not circumstance-dependent in the way in which its extension is. The content of an expression is *what* we evaluate in or at a circumstance; hence it is determined prior to the encounter with the circumstance (Kaplan 1989). For example, the content of the description 'the President' (or 'the President of the US', or 'the President of the US in 1999') can be construed as a function from situations to individuals. Once an appropriate circumstance is given, the function determines the extension of the description in that circumstance, that is, the individual who is the value of the function. The value of the function is circumstance-dependent, but the function itself is not.

In meta-representational contexts, however, substituting an expression for another one with the same content possibly affects truth-value. 'Ophthalmologist' and 'eye-doctor' are synonymous, hence they have the same content (a function from situations to sets of individuals). It follows that the content of a sentence in which one expression occurs will remain the same after substituting the other expression for it. That will be so whether or not a circumstantial shift occurs, since the content of an expression is not circumstance-dependent. Now if the content of the sentence remains the same, its truth-value also will remain the same. Despite all this, it is intuitively evident that substituting 'ophthalmologist' for 'eye-doctor' in (1) can change truth-value: it is possible for John to believe that Peter is an eye-doctor and to disbelieve that he is an ophthalmologist. In other words, (1) and (2) can be both true.

(1) John believes that Peter is an eye-doctor

(2) John does not believe that Peter is an ophthalmologist

To account for that puzzling fact, we must follow Quine and assume that in (2) the word 'ophthalmologist' is somehow *mentioned*, as if the speaker had said

(2*) John does not believe Peter to be an 'ophthalmologist'

This quotational element makes the meta-representation opaque. Yet I maintain that meta-representations as such are fundamentally transparent. They rest on a form of simulation which entails that the ascriber and the ascribee are in the same boat. What happens in (2*) is that *another form of simulation takes place*, which accounts for the possibility of conceptual or ontological divergences, and for intensional substitution failures more generally. That other form of simulation is specifically communicational, and it is closely tied to the phenomenon of quotation.

6. Context-shifting

When we quote someone else's words, arguably, we engage in a form of play-acting: we simulate the person in question by actually making the utterance we're ascribing to her. Herb Clark makes that point vivid by means of an example in which someone, quoting Greta Garbo, utters 'I want to be alone' in a Swedish accent while clutching his arms to his chest in a Garboesque pose (Clark 1996: 175). The quoter in that example is obviously playing the part of Garbo. According to the simulation theory of quotation (Ducrot, Clark, Wierzbicka), something similar takes place whenever we quote someone. Quotation involves a context-shift: the quoter simulates the person whose speech or thought he is reporting, much as an actor simulates the character whose part he is playing, or as the novelist simulates the narrator. He temporarily pretends to be that person. That account goes a long way toward explaining why 'I', in quotations, does not refer to the person who quotes, but to the quoted person.

If this is right, opacity examples like (2*) are similar to the following examples, where, in the course of reporting the ascribee's speech act, the speaker mimics him or her by phrasing and/or pronouncing the complement sentence in a certain way:

> To which Mr Bailey modestly replied that he hoped he knowed wot o'clock it wos in gineral.
> (Dickens, *Martin Chuzzlewit*, cited in Clark and Gerrig 190: 791)

> Une vieille femme... vint au seuil et me demanda *qué que j'voulais*, d'une voix traînante et hargneuse. (Barbey d'Aurevilly, *L'Ensorcelée*)

At the same time as he reports the ascribee's talk, using indirect speech in the normal way, the speaker *shows* what that talk was like.

Whenever intensional substitution failures occur in meta-representations, that is evidence that the words in the complement sentence are used *echoically*. The ascriber does not merely replicate the content of the object-representation, he also mimics the ascribee by actually expressing that content the way the ascribee herself would. If we substitute other words with the same content, the echo is affected, hence the overall meaning of the utterance changes. This accounts for failures of substitutivity.

In the examples I have given, the words, although used echoically, keep their normal semantic values. The echo simply adds a further layer of meaning to the basic content of the utterance. That is a very common phenomenon. Often one uses a word while at the same time implicitly ascribing that use to some other person (or group of persons) whose usage one is blatantly echoing or mimicking. Thus one might say:

That boy is really 'smart'

In such examples one is quoting, but at the same time using the words with their normal semantic values. Sometimes, however, the contextual shift affects the content of what is said: the words are taken not with their ordinary meanings (the meanings they have in the actual context), but with the meanings they have in the shifted context, i.e. for the person one is mimicking. Thus I can refer to some object, A, using the name of another object, B, in quotes, providing the person I am mimicking uses the name for B as a name for A. I may well say

'Quine' has not finished writing his paper

and refer, by the name 'Quine' in quotes, not to Quine but to that person whom our friend James mistakenly identified as Quine the other day. Any word can, by being quoted in this echoic manner, be ascribed a semantic value which is not its normal semantic value, but rather what some other person takes to be its semantic value. For example, if James has a poor mastery of English and uses 'paper' in lieu of 'poster', I can ironically use (1) to mean (2):

(1) Quine has not finished writing his 'paper'

(2) Quine has not finished writing his poster

I can also use a non-word provided it belongs to the lexicon of the person whose usage I am mimicking, as in (3):

(3) My three-year-old son believes that I am a 'philtosopher'

In the cases I have just mentioned a context-shift takes place but it is *translinguistic*. It takes us from the actual context (in which the normal rules of English apply) to a pretend context in which deviant semantic rules apply. In a closely related type of case, the linguistic norms are on the other side: the speaker (or thinker) does not fully master the rules of the public language but defers to some authority who does. In Burge's most famous example (Burge 1979), someone complains that she has arthritis in her thigh. Since arthritis is an inflammation in the joints, that cannot be the case. But the person who so complains does not know the exact meaning of the word 'arthritis' – she has only a partial and incorrect idea. That does not prevent her from using the word deferentially. In virtue of what Putnam has called 'the division of linguistic labor' (Putnam 1975), she succeeds in expressing the proposition that she has arthritis in the thigh – a proposition which, of course, cannot be true.

Instead of ironically mimicking a deviant use, as in (1) and (3), the speaker in Burge's example mimics the use of those who, unlike her, do master the concept. Despite the differences, I take it as intuitively evident that the same sort of context-shift is involved in both cases. In both cases the speaker uses someone else's language. Deference is a form of simulation which is made possible by the existence of a public language. Thanks to the deferential mechanism, we can use public symbols to express the contents which those symbols have for other people in our community. In this way we can even express concepts which we do not possess. Conceptual and ontological heterogeneity is thus made possible.

Suppose Mary says:

(4) My children believe that Santa Claus can't come until we're all asleep

Since Santa Claus does not exist, Mary cannot really assert that her children believe *him* to have such and such properties: she cannot ascribe her children a singular belief concerning Santa Claus, for that would commit her to the existence of Santa Claus. This is a straightforward application of the Inheritance Principle: if the internal sentence is truth-valueless because Santa Claus does not exist, then the complex sentence (4) is bound to be truth-valueless too. But what Mary can do, and what she does, is engage into pretense and shift the context: she temporarily pretends that Santa Claus exists, and within that pretense ascribes her children the belief that *he* can't come until they are all asleep.

This analysis of statements like (4) was first put forward by John McDowell twenty years ago, in the following passage:

> Suppose an interpreter finds an expression – say, 'Mumbo-Jumbo' – which functions, syntactically, like other expressions which he can construe as names, but for which he can find no bearer, and reasonably believes there is no bearer. (...) A sincere assertive utterance of a sentence containing a name with a bearer can be understood as expressing a belief correctly describable as a belief, concerning the bearer, that it satisfies some specified condition. If the name has no bearer (in the interpreter's view), he cannot describe any suitably related belief in that transparent style. (...) In practice, an interpreter might say things like 'This man is saying that Mumbo-Jumbo brings thunder', and might explain an utterance which he described that way as expressing the belief that Mumbo-Jumbo brings thunder. That is no real objection. *Such an interpreter is simply playing along with his deluded subject – putting things his way*.
>
> (McDowell 1977: 124–127; emphasis mine)

According to McDowell, the interpreter's assertion, to the effect that the subject says/believes that Mumbo-Jumbo brings thunder, is a pretend assertion. In virtue of the Inheritance Principle, such an assertion presupposes, and can be made seriously only if the speaker believes, that Mumbo-Jumbo exists. If the interpreter takes the name 'Mumbo Jumbo' to have no bearer, he must either refrain from reporting the belief in the same style in which it is expressed (that is, using a simple subject-predicate sentence: 'Mumbo-Jumbo brings thunder'), or, 'playing along with his deluded subject', he must temporarily pretend that Mumbo Jumbo exists. Through that pretense the context is shifted. In the actual context of use the speaker does not believe (let alone presuppose) that Mumbo Jumbo exists; interpreted with respect to the actual context, the utterance would therefore violate the principle that "if a sentence *x* semantically presupposes a proposition φ..., then φ is presupposed by the speaker" (Stalnaker 1978: 326). In interpreting the utterance we are therefore led to shift the context and interpret the utterance with respect to some imaginary context in which the speaker does believe in Mumbo Jumbo. That mechanism of context-shift corresponds to Stalnaker's description: From "an apparent violation of the rule... one may conclude that the context is not as it seems" (*ibid.*).

7. Conclusion

In this talk three grades of simulative involvement have been distinguished in belief reports and other meta-representations. If we know what someone believes and/or desires, we can conduct inferences by 'assuming' those beliefs and desires and seeing what results. But simulation is not merely useful for

inferentially *exploiting* meta-representations; there is a sense in which it is involved also in the very process of *entertaining* a meta-representation. That is so because genuine meta-representations include the object-representation as a proper part: thus the representation that John believes that it is raining contains the representation that it is raining and cannot be entertained unless the latter also is entertained. It follows that we cannot ascribe contents to a radically alien creature. We can only ascribe contents to creatures which are sufficiently like us and whose conceptual resources we share.

The phenomenon of opacity raises an obvious problem for this view. There may be all sorts of conceptual divergence between ascriber and ascribee. Thus a belief report may be true even though the content sentence contains what the ascriber takes to be an empty name ('My son believes that Santa Claus will come tonight'). To account for such facts, and for 'hyperintensionality' more generally, I argued that meta-representations involve yet another form of simulation. That third form of simulation is specifically communicational, and it is closely tied to the phenomenon of 'deference'. I suggested a treatment of it in terms of 'context-shift' – a treatment which I pursue in my book *Oratio Obliqua, Oratio Recta*.

Notes

1. See Fodor (1983) on 'belief fixation'.

2. As Stalnaker puts it, 'to accept a proposition is to act, in certain respects, as if one believed it' (Stalnaker 1987: 80).

3. This field of study, they say, 'appears exceptionally rich and promising' (Kahneman & Tversky 1982: 204).

4. In Dummett's terminology, a phrase like 'Leibniz's Law', which picks out a thought, "does not *express* that thought, since it is possible to understand the phrase without knowing which thought it picks out" (Dummett 1993: 173). Dummett does not mention 'that'-clauses in that passage, but I don't see how he could deny that a 'that'-clause expresses the thought it picks out. Mark Richard similarly argues that 'that'-clauses *articulate* the propositions they refer to, while phrases like 'Leibniz's Law' do not (Richard 1993).

5. Bühler distinguishes two sorts of simulative presentification: either the subject mentally experiences the imagined situation as if that situation was actually present around him – as if it was (a part of) the hic-et-nunc situation; or he experiences the imagined situation as if his egocentric coordinates were displaced to that situation – as if he himself was in that situation:

> To put it in the manner of a parable, either Mohammed goes to the mountain or the mountain goes to Mohammed... What is imagined, especially when movable things

> such as people are concerned, often comes to us, that is, into the given order of actual perception, within which it can be localized, though not quite "seen"... The imagined thing that appears to the mind's eye... can receive a place in front of, next to or behind me, located directly among the things in the room in which I am, among the things that I in part perceive, in part imagine... The exact opposite occurs in the second type, in which Mohammed goes to the mountain... One is displaced in imagination abruptly, suddenly to the geographical place of what is imagined, one sees what is imagined in front of one's mind's eye from a certain reception point which one can identify and at which one is situated in imagination. If one turns around in imagination, one sees what was behind one's back, if one walks on, one sees the things again in imagination as one once saw them when really walking there. (Bühler 1934: 150–151)

Bühler mentions various experiments in support of this distinction and at several points refers to Jakob Segal's in-depth study of mental simulation (Segal 1916).

References

Bühler, K. (1934). *Theory of Language*. (D. F. Goodwin, Eng. trans.) Amsterdam: John Benjamins. (Orginal work published 1934.)

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy, 4*, 73–121.

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. & Gerrig, R. (1990). Quotations as demonstrations. *Language, 66*, 764–805.

Currie, G. (1990). *The Nature of Fiction*. Cambridge: Cambridge University Press.

Currie, G. (1995). Visual Imagery and the Simulation of Vision. *Mind and Language, 10*, 25–44.

Dinsmore, J. (1991). *Partitioned Representations*. Dordrecht: Kluwer.

Dretske, F. (1988). *Explaining Behavior*. Cambridge, MA: MIT Press.

Ducrot, O. (1984). Esquisse d'une théorie polyphonique de l'énonciation. In O. Ducrot, *Le Dire et le Dit* (pp. 171–233). Paris: Minuit.

Dummett, M. (1993). *The Seas of Language*. Oxford: Oxford University Press.

Fauconnier, G. (1985). *Mental Spaces*. Cambridge, MA: MIT Press.

Fodor, J. A. (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

Geach, P. (1957). *Mental Acts: their Content and their Objects.* London: Routledge and Kegan Paul.

Heal, J. (1998). Co-Cognition and Off-Line Simulation: Two Ways of Understanding the Simulation Approach. *Mind and Language, 13*, 477–498.

Kahneman, D. & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 201–207). Cambridge: Cambridge University Press.

Kaplan, D. (1989). Demonstratives. In J. Almog, H. Wettstein, & J. Perry (Eds.), *Themes from Kaplan* (pp. 481–563). New York: Oxford University Press.

Lewis, D. (1978). Truth in Fiction. Reprinted (with postscripts) in his (1983). *Philosophical Papers*, vol. 1 (pp. 261–280). New York: Oxford University Press.

McCarthy, J. (1993). Notes on Formalizing Context. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, vol. 1 (pp. 555–560). San Mateo: Morgan Kaufmann Publisher.

McDowell, J. (1977). On the Sense and Reference of a Proper Name. Reprinted in A. W. Moore (Ed.) (1993). *Meaning and Reference* (pp. 111–136). Oxford: Clarendon Press.

Nichols, S. et al. (1996). Varieties of Off-Line Simulation. In P. Carruthers & P. Smith (Eds.), *Theories of Theories of Mind* (pp. 39–74). Cambridge: Cambridge University Press.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Putnam, H. (1975). The Meaning of Meaning. In K. Gunderson (Ed.), *Language, Mind, and Knowledge* (pp. 131–193). Minneapolis: University of Minnesota Press.

Quine, W. V. O. (1995). Reactions. In P. Leonardi & M. Santambrogio (Eds.), *On Quine* (pp. 347–361). Cambridge: Cambridge University Press.

Recanati, F. (2000). *Oratio Obliqua, Oratio Recta: An Essay on Meta-representation*. Cambridge, MA: MIT Press.

Richard, M. (1993). Articulated Terms. *Philosophical Perspectives, 7*, 207–230.

Sechehaye, A. (1926). *Essai sur la structure logique de la phrase*. Paris: Champion.

Segal, J. (1916). *Uber das Vorstellen von Objekten und Situationen: Ein Beitrag zur Psychologie der Phantasie*. Stuttgart: Spemann.

Sperber, D. (1985). Apparently Irrational Beliefs. *On Anthropological Knowledge* (Chapter 2). Cambridge: Cambridge University Press.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics, 9*, 315–332.

Stalnaker, R. (1987). *Inquiry*. Cambridge, MA: MIT Press.

Walton, K. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts.* Cambridge, MA: Harvard University Press.

Wierzbicka, A. (1974). The Semantics of Direct and Indirect Discourse. *Papers in Linguistics, 7*, 267–307.

Reply to François Recanati

Jérôme Pelletier
Institut Jean-Nicod, CNRS, Paris

1. Simulation and assumption

François Recanati claims that assumption is simulative both in a technical and in a more phenomenological sense. Assumption is simulative in a technical sense since it uses the inference mechanism off-line and it is simulative in the phenomenological sense in so far as it is often accompanied by a certain kind of *as if* behaviour: 'when someone assumes that *p*, he or she behaves *as if* the representation that *p* was validated'.

Let's reconstruct Recanati's reasoning:

1. Assumption is a kind of acceptance
2. Acceptance is the disposition to exploit a representation, i.e. to feed a representation to the inferential device. It is a disposition to run the inferential device within the belief module.
3. The function of the inferential device is to exploit representations which are validated. Acceptance, *qua* disposition to run the inferential device, exists only because of that function.
4. There is a divergence between the natural function of the inferential device and the way the inferential device is used in assumption, that is 'off-line'.
5. There is simulation whenever a mechanism which has a function within a system is used 'off-line' (and sometimes when there is a certain kind of *as if* behaviour).
6. Therefore assumption is simulative.

I am reluctant to accept Recanati's conclusion: I would not be ready, as Recanati does, for instance to describe the children who pretend that a certain stick is a sword as 'assuming' that a stick is a sword. One assumes a certain proposition either in a polemical or in a heuristic context. Moreover assump-

tion can be described as an adult purposeful activity and this is rarely the case with children's games of make-believe. I believe that Recanati's conclusion goes too much against our colloquial use of the term 'assumption'.

At the same time, I accept, with Recanati, premiss 1 and premiss 5. So my disagreement with him lies in 2–4. The main idea which I reject in 2–4 is that there would be something like *the* unique natural function of acceptance. I see acceptance as a primitive mental state of regarding a certain proposition *as* true, but a mental state which has no privileged function within the belief module. Acceptance is 'designed' to work in various cognitive states as different as imagination and assumption on one side and belief on the other side. On that different basis, assumption is, like belief, a case of acceptance (premiss 1) since it is a case of regarding a proposition *as* true. But it is no more needed to introduce a simulative ingredient to specify the distinction between belief and assumption. Let's draw a different picture from Recanati's of the relations between belief, acceptance and assumption.

What distinguishes belief and assumption is mainly their different aims, their different ways of accepting a proposition, not the fact that, in assumption, acceptance would be running off-line. Assumption is the mental state it is because it is a state of acceptance of a proposition which does not take into account the *real* truth-value of the proposition accepted. In order to assume a proposition one needs only regard that proposition *as* true without considering the *real* truth-value of the proposition. By contrast, a belief is a way of accepting a proposition or of regarding a proposition as true while at the same time considering that it is its *real* truth-value. I cannot believe a proposition without believing it, maybe wrongly, to be *really* true whereas in assumption I only accept a proposition, that is I regard it as true in a quite different spirit, for the sake of the argument (cf. Velleman 1992). So my main disagreement with Recanati lies with the idea expressed in 2–4 that there would be something like the *real* or *natural* function of acceptance and which would be to feed the inferential sub-system of the belief faculty.

2. Simulation and prediction

Recanati's second thesis is that reasoning on someone else's beliefs is a kind of simulative reasoning. When one reasons on someone else's beliefs, the exercise of 'projection', 'retrojection' and 'deduction' makes it possible to predict someone else's beliefs. And this attempt to reason within a person belief's space is a kind of 'simulative reasoning'.

The simulative part of the so-called 'simulative reasoning' is the operation of 'cognitive projection', that is the process of assuming or accepting the propositions believed by someone else, what Recanati calls 'the object representation'. Simulation here is understood as the process of running off-line some version of someone else's beliefs. In order to calculate someone else's mind, through cognitive projection, one is able to remove the prefix 'John believes that. . .' and takes on temporarily John's belief.

Are we really ready to admit that somebody who reasons on someone else's beliefs 'takes on' this person's beliefs or imagines believing *herself* such and such propositions? I doubt that the process of belief ascription requires such participation on the side of the ascriber. But my point might be simply terminological since Recanati claims that the operation of cognitive projection which lies at the heart of the process consists in 'assuming' or 'accepting' the propositions believed by someone else. If then one understands 'assumption' and 'acceptance' the way I did in the preceding section, then I totally agree with Recanati's description of the phenomenon of belief ascription and prediction.

Recanati tries then to resist a counter-example levelled against this simulative approach to reasoning, a counter-example which shows the limits of this method of reasoning. This method of reasoning can be used only if the premises already belong to the belief space of the person whose beliefs one tries to predict. No premises external to the 'assumptive space' can be used in such reasoning.

The fact that deductions on the basis of projections are limited to the premises already belonging to the belief space of the believer raises a difficulty previously noticed by Proust in her paper in this volume (201–228). As a matter of fact, there is here a risk of circularity: in order to reason simulatively within someone else's belief space, the ascriber presumably needs to know the premises authorised in the reasoning, which means that she already needs to know the content of the belief space of the ascribee. But this is precisely why the simulative reasoning was started in the first place.

Moreover I wonder whether this method for 'exploiting' meta-representations, especially the process of retrojection, is really faithful to the logic of belief. The fact is that beliefs do not seem closed under deduction. We are not perfectly logical agents, in particular, we do not believe all the consequences of our beliefs, for instance, we do not believe the contradictions that might follow from our beliefs (cf. Dretske 1970). But simulatively reasoning on someone else's beliefs might in some cases lead to ascribe to someone the contradictions he or she does *not* believe in. And in that case, one would loose all the in-

terest of this kind of reasoning which is precisely to help us predict someone else's beliefs.

3. Simulation and meta-representation

I come to the main part of Recanati's paper. According to Recanati, meta-representations are 'intrinsically simulative'. One needs to put oneself into someone else's shoes, one has to think what she thinks, in order to entertain a genuine meta-representation. Whereas in simulative reasoning, the idea was that in order to *calculate* someone else's belief, one has to take on the contents believed by that person, here the idea is that in order to *represent* someone else's beliefs, in order to construct a meta-representation, one has first to entertain the contents represented by that person.

3.1 The iconicity of meta-representations

What is really new in this proposal? The followers of Gordon and Goldman know very well that simulation is centrally involved in attributing mental states to others. But Recanati does not just hold this view. He presents new evidence in favour of this view based on the linguistic format of our belief reports. According to Recanati, the simulative essence of meta-representations follows from the intuitive observation that in order to meta-represent, to have a second level belief about a first level belief, one needs to *display* the content believed at the first level. This is what Recanati calls the 'iconic' dimension of meta-representations. Meta-representations are 'iconic' in so far as they resemble or replicate the beliefs they are about.

Recanati is right to insist that in belief reporting some *picturing* takes place. And Recanati is careful enough to speak here vaguely of resemblance since there are many different ways of *displaying* a content believed at the first level, different ways to think what another person thinks. Imagine for instance that:

(1) He is a jolly good fellow

is sincerely uttered by Joëlle while demonstrating Pierre. It seems that I could truly report her belief by uttering any of the following sentences:

(2) Joëlle believes that Pierre is a jolly good fellow

(3) Joëlle believes that he is a jolly good fellow [demonstrating Pierre]

(4) Joëlle believes that you are a jolly good fellow [addressing Pierre]

(2) and (4) are true belief reports which display the content believed by Joëlle in a semantically faithful way even though they are not syntactically faithful to the way Joëlle expressed her belief as in (3). (2) and (4) 'resemble' less than (3) Joëlle's way of expressing her belief.

When I report Joëlle's belief the way I do in (2), this is because I feel free to exploit the fact that I know the proper name of the person demonstrated by Joëlle even though I know that Joëlle ignores his name. In these circumstances, while truly attributing to Joëlle the belief that Pierre is a jolly good fellow, I would be *displaying* the content of Joëlle's belief but in a way different from her. What Recanati calls the 'object representation' in a meta-representation might then be in some cases a reconstruction of the ascribee's thought from the ascriber's perspective.

Recanati claims that "Tom stated that identical objects have all their properties in common" is a genuine meta-representation contrary to "Tom stated Leibniz's law" since the content of Tom's statement is displayed only in (2). But the problem is that, since many meta-representations rely as in (2) and (4) on a reconstruction on the side of the ascriber, in many intermediate cases one might not be able to tell whether or not one is dealing with what Recanati calls a 'genuine meta-representation'.

3.2 The transparency thesis

The 'iconic' dimension of meta-representations in turn explains, according to Recanati, why the ascriber needs first to *entertain* the content of the first-level representation he or she attributes. This last move leads Recanati to claim that meta-representations are 'transparent' representations. Meta-representations do not constitute an opaque interface between the ascriber and the ascribee's thoughts since the ascriber needs to entertain the semantic content of the representation he attributes to the other person.

Whenever a meta-representation displays the content x of an object-representation, then the meta-representation is bound to be about x. Let's quote Recanati: "a genuine meta-representation dS (where d is the tag and S the radical) satisfies the following schema:

> Schema (I):
> One cannot entertain the proposition that dS without entertaining the proposition that S.

For example:

> One cannot entertain the proposition that John believes that grass is green without entertaining the proposition that grass is green.

According to Recanati, the linguistic format of our belief reports, the way belief reports are displayed, would then be another evidence *via* the 'transparency thesis' in favour of the simulation theory of meta-representations.

Recanati's transparency thesis has for it a pre-philosophical intuition about the truth of our belief reports. According to that intuition, a true belief report contains a that clause whose terms have the same *references* as the ones used by the subject of the report. As is well-known, this pre-philosophical intuition goes against what has now become the standard Fregean philosophical intuition according to which the that-clause of a belief report has to express some sort of conceptual content that the subject of the report believes in order for the report to be true. To use Davidson's phrase, Recanati tries with Schema (I) to recover this 'pre-Fregean semantic innocence' (Davidson 1968: 108).

Schema (I) is liable to face several immediate difficulties which Recanati tries to deal with in the rest of the paper. I will mention one which is apparently omitted in Recanati's paper. What happens to Schema (I) when the proposition 'believed' by John is an 'impossible proposition'? Let's suppose for the sake of the argument that there is such a thing as an 'impossible proposition'. Does one need to entertain the proposition that $2 + 2 = 5$ or the proposition that Hesperus is not Phosphorus in order to entertain the proposition that John believes that $2 + 2 = 5$ or the proposition that John believes that Hesperus is not Phosphorus? Imagine that John is experiencing the Capgras delusion and believes that, Mary, his spouse has been replaced by an impostor who looks just like her. When the doctor commenting on Johns' situation says:

(5) John believes that Mary is not Mary,

must we accept with Recanati that the doctor entertains the thought that Mary is not Mary? If the entertaining of the object proposition is a necessary condition of the entertaining of the meta-proposition, if, as Recanati claims, there cannot be an entertaining of the meta-proposition without an entertaining of the object proposition, each time one attributes a belief in an 'impossible proposition', one would then entertain an 'impossible proposition' which is a difficulty for Schema (I).

Moreover, in these cases of belief with a logically or metaphysically impossible content, it might well be that the correct account of John's state of mind is that John *claims* to believe an impossible proposition or that John *says* that he

believes an impossible proposition. So why should this kind of report impose on the ascriber to entertain himself or herself impossible propositions? In order to give this kind of report, the striking fact is that it seems necessary *not* to put oneself into John's shoes, *not* to think or replicate in thought what he apparently thinks. And these intuitions go against Schema (I). The main purpose of a belief report with an impossible content is neither to think nor to talk 'transparently' about the impossible *states of affairs* represented by John but only to talk of John's impossible *mental representations* as such without entertaining the content of John's representations.

In the rest of the paper, Recanati develops a thesis concerning the decoupling procedure at work each time there is a primary representation and a tag, and he tries to defend the 'transparency thesis' expressed in Schema (I) against various objections. I deal with these claims in the following sections.

4. Meta-representational prefixes vs. 'scene-setting' prefixes

Recanati claims that meta-representations are a special case of the decoupling procedure used in many other circumstances, that is whenever one finds a primary representation and a tag or prefix. For instance, when we have meta-representational prefixes such as 'In the picture, grass is green' and 'In the film, grass is green', or spatial prefixes such as 'In Chicago, grass is green', according to Recanati what we really have is an attempt by the speaker to go imaginatively beyond his egocentric situation and to describe a distinct situation. This is something previously noticed by Perner:

> When exclaiming 'Look, there is Judy in the picture. She is wearing blue' (. . .). With the word 'Judy', we make a direct reference to Judy being in the picture, and the expression 'in the picture' serves as a context-marker in much the same way that, in the statement 'Yesterday at your party Judy was wearing blue', the expression 'Yesterday at your party' marks a difference in the spatio-temporal context which differentiates it from that of current reality.
>
> (Perner 1993: 129)

Recanati's proposal is to interpret the presence of these various tags or operators as an invitation to simulation. But shouldn't we distinguish the simulative role played by true meta-representational prefixes such as 'In the film' and 'According to John' from the non-simulative role played by non-meta-representational prefixes or 'scene-setting prefixes' such as 'In Chicago' or 'Yesterday'?

As a matter of fact, I think that Recanati's proposal might be plausible if it were restricted to meta-representational operators such as 'In the fiction, . . .' 'In the picture. . .' or 'In the film. . .'. Understanding some remarks made with these operators may require to continue the simulation initiated by the author of the fictional story, picture or film. This is actually Evans (1982) and Walton's (1990) thesis on discourse about fiction or 'metafictive' discourse.

But even in these cases, I find the interpretation of metafictive discourse given by van Inwagen (1977) or Currie (1990) more plausible. Let's consider the following 'metafictive' utterance:

(6) In Shakespeare plays, some characters are neurotic.

The tag 'In Shakespeare plays' is an invitation to adopt a serious attitude towards the plays, even a scientific one. This is an invitation to decompose the play in its structure, to see for instance that plays contain among other things characters and that these characters have certain properties. I do not see the tag here as an invitation to a simulation. There is no simulation or decoupling procedure when one talks of the play in this case.

But even though one endorses the Evans-Walton-Recanati thesis on metafictive discourse and tags, would it be possible, following Recanati, to generalise it to *non*-meta-representational or to what may be called 'scene-setting' prefixes? Recanati is ready to generalise the simulative analysis to many prefixes which have nothing to do with meta-representational matters. According to Recanati, 'In Chicago, it is raining' is to be understood along the same lines as 'In the film, it is raining', that is as involving cognitive projection or simulation. It is true that there is in 'In Chicago, it is raining' a decoupling in so far as the object representation is not tokened to describe the environment of the utterer. But this decoupling has nothing to do with a simulation. I would not say like Recanati that in this case "one simulatively entertains a representation decoupled from the egocentric situation" or that one "simulates perception". If one really thinks that simulation or imagination is involved in the Chicago statement – something which I find doubtful since the Chicago statement is the expression of a judgement – a more accurate description of the situation would be that in uttering 'in Chicago, it is raining', one is led to imagine that something happens in Chicago – a case of *impersonal* imagining according to Currie (1995: 166) – without putting oneself into anybody's shoes, in particular without *simulatively* imagining observing or seeing the rain falling in Chicago. I do not see the need to locate in imagination the rain event in relation to myself when I judge that it is raining in Chicago.

As an aside, I add that I also wonder whether Recanati's theory applies to statements such as 'Yesterday, it was raining'. With temporal prefixes or tags such as 'yesterday' or 'tomorrow', primary representations being already at the past or future tense and, by the same token, being already decoupled from the present situation of the utterer, it is difficult to understand the presence of the tag as inducing an extra decoupling.

5. Substitution failures without opacity?

At the end of the paper, Recanati tries to resist various counter examples to the 'transparency thesis' concerning genuine meta-representations. All counterexamples are opacity examples, that is examples of meta-representational sentences wherein substitution of co-referring terms in the object proposition does not preserve truth value.

Recanati's difficult strategy is to maintain that all these meta-representational sentences are true cases of substitutional failures and that these meta-representational sentences are nonetheless transparent.

As I understand him, Recanati denies that 'believes' and like verbs *really* create opaque contexts. But still Recanati claims that in belief reports, substitution of co-referring terms fails.

One traditional line of argument in favour of the thesis that 'believes' and like verbs do not *really* create opaque contexts is to claim that the appearance of opacity in meta-representational sentences has arisen from a confusion between what a statement says and what it implicates. Salmon (1986) has defended such a strategy and claims that the apparent differences in truth value between:

(7) Hammurabi believes that Hesperus is visible in the evening

(8) Hammurabi believes that Phosphorus is visible in the evening

results from a confusion between what might be *pragmatically* conveyed by such sentences – that is a difference in truth value- and what these sentences *semantically* do express – that is the same proposition.

Recanati does not follow such a strategy. The difference in truth-value between (7) and (8) is, for Recanati *contra* Salmon, not just *pragmatically* communicated. Substituting 'Phosphorus' for 'Hesperus' in (7) could really lead to a change in truth-value. But, according to Recanati, this real change should not lead the semanticist to conclude that belief reports are opaque. Rather, the

semanticist should look more closely at the ascriber's way of using the language when reporting a belief. She will then find that words in the ascriber's mouth are used either in a quotational or in a deferential way, that is in ways which all involve simulation. This is the reason why the opacity conclusion is not imposed to the semanticist and the transparency thesis is saved. Let's try to reconstruct Recanati's argument:

1. Meta-representational sentences are true cases of substitution failures. Coreferential singular terms in belief reports are not substituable *salva veritate*.
2. The best explanation of substitution failures in meta-representational sentences is either that there is a circumstancial shift or that the ascriber implicitly uses the language in a quotation or in a deferential way.
3. Leaving aside circumstancial shifts, quotational and deferential uses of language involve context-shift and simulation.
4. Terms used in the embedded portion of a belief sentence contribute their normal semantic value even though they are not substituable with co-referring terms which is the case when they are used in a quotational or deferential way.
5. Conclusion: substitution failures are compatible with transparency. The traditional opacity examples are in fact examples of quotational or deferential uses of language.

Recanati claims in premise 3 that some quotational use of language involves context-shift and simulation. In some particular cases such as:

(9) 'Quine' has not finished writing his paper

the best explanation of the presence of quotation marks is that the intention of the speaker is not to use the proper name inside the quotation marks with its normal semantic value but with the semantic value that another person attributes to it. When reporting James's false belief, I might say:

(10) James believes that 'Quine' has not finished writing his paper.

In this report, I would be mimicking my friend James's way of using the name 'Quine' to refer to a different person from the late philosopher. This quotational use of language explains why one cannot substitute *salva veritate* corerefential terms with 'Quine' in (10).

Following Putnam and Burge, Recanati adds in premise 3 that in many cases speakers who have only a partial mastery of the rules of the public language mimic the use of those who master the language. Deference is then con-

strued by Recanati as another form of simulation. In both cases, using words inside quotation or deferentially is construed as a way of using someone else's language within one's language or as playing the part of someone else.

Now it is true that *one way* of depicting a person's mental life is to include in one's belief report the very words the person used, either in a quotational or deferential way. An appeal to these special uses of language can help to explain *some* cases of substitutional failures in belief reports. And this is the major claim made by premise 2. But the main problem with Recanati's argument is that it seems *a priori* very difficult to account for *all* cases of susbstitution failures on the basis of premise 2, for instance to account for Kripke's Pierre puzzling beliefs by appealing to a simulative use of language.

In the end of the paper, Recanati gives his own analysis of what is traditionally considered as an opaque statement. Recanati analyses:

(11) My son believes that Santa Claus will come tonight

in Walton's way that is as involving simulation or pretense. (11) is according to Recanati a pretend assertion. The mother saying (11) engages into pretense and shifts the context: she temporarily pretends that Santa Claus exists, and within that pretense ascribes her children the belief that *he* will come tonight. The transparency is saved, according to Recanati, thanks to the simulation of the ascriber: since the mother pretends that there exists a certain person who is called 'Santa Claus' and who will come tonight, the existential generalisation from (11) can be inferred within the pretense.

Remember that what Recanati means by 'transparency' is that

> "whichever state of affairs the object-representation represents, the meta-representation also represents".

In this particular example, the object representation 'Santa Claus will come tonight' does not represent any state of affairs since Santa Claus does not exist. Therefore, in this analysis, neither does the meta-representation represent any state of affairs except within a pretense. The ascriber and the ascribee are in the same *fictional* boat.

Pace Recanati, I have the impression that the meta-representation, *contrary to the primary representation*, really does represent a state of affairs and does not only pretend to represent something. Belief ascriptions such as (11) are not pretend ascriptions, whatever the status of the primary representation might be. In this particular case, I do not see why the mother would pretend to assert (11) since she really believes that her son believes that Santa Claus will come tonight. The mother does not imagine that her son believes that Santa Claus

will come tonight. A pretend assertion reflects at the linguistic level what a pretend belief or imagined belief is at the mental level. But here, there is no reason to postulate a pretend belief on the side of the mother. Actually what the mother could even say is that her son *claims* to believe that Santa Claus will come tonight. She would then state a fact about the actual world without pretending anything. And this is something that Recanati's account might have difficulties to explain.

References

Davidson, D. (1968). On Saying That. Reprinted in his (1984). *Inquiries into Truth and Interpretation* (93–108). Oxford: Clarendon Press.

Dretske, F. (1970). Epistemic Operators. *Journal of Philosophy, 67*, 1007–1023.

Evans, G. (1982). *The Varieties of Reference*. In J. McDowell (Ed.). New York: Oxford University Press.

Currie, G. (1990). *The Nature of Fiction*. Cambridge: Cambridge University Press.

Currie, G. (1995). *Image and Mind – Film, Philosophy and Cognitive Science*. Cambridge: Cambridge University Press.

Perner, J (1993). Rethinking the meta-representation theory. In S. Baron-Cohen, H. Tager-Flusberg & D. J. Cohen (Eds.), *Understanding other Minds: Perpsectives from Autism* (pp. 112–137). Oxford: Oxford University Press.

Proust, J. (2002). Can "radical" simulation theories explain psychological concept acquisition? In J. Dokic & J. Proust (Eds.), *Simulation and Knowledge of Action* (pp. 201–228). Amsterdam: John Benjamins.

Salmon, N. (1986). *Frege's Puzzle*. Cambridge: MIT Press.

Van Inwagen, P. (1977). Creatures of Fiction. *American Philosophical Quaterly, 4*, 299–308.

Velleman, D. (1992). The Guise of the Good. *Noûs* 26, 3–26. Reprinted in Velleman, D. (2000). *The Possibility of Practical Reason* (pp. 99–122). Oxford: Oxford University Press.

Walton, K. (1990). *Mimesis as Make-Believe-On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.

Chapter 7

Mental simulation, dialogical processing and the syndrome of autism

Donald M. Peterson
London Knowledge Lab./Institute of Education, University of London

1. Introduction

Several authors, from Bartlett (1932) and Piaget (1937/1955) onwards, have argued that the human cognitive system uses a schema-based architecture. That is: our knowledge is not stored in a flat manner, but is organised into more or less self-contained units which serve as ready formulae for performing particular tasks. Investigation in artificial intelligence has further developed the notion, as in work by Minsky (1975) on 'frames', Shank and Abelson (1977) on 'scripts', and Arbib (1989) and others on 'schemas'. One of the main motives for this proposal is that a schema (composed of rules, data and goals) serves to avoid the labour of extensive or complete search through the system's knowledge-base: faced with a problem of a particular type, the system indexes an appropriate schema rather than searching throughout its knowledge-base in order to decide what to do. And since, in a system of any realistic size, search may become impracticably complex, some device is required to avoid this: using 'brute-force' search, for example, complexity of search will grow exponentially with problem-size (see Korf 1988), thus becoming unmanageable.

A schema-based architecture, however, creates what might be called the *problem of context*. The limited and self-contained nature of a single schema is a two-sided coin: it avoids the need for unmanageable search, but it is unlikely to contain code for all eventualities. That is, there may exist a circumstance or context which is represented outside a schema but which is relevant to the control of mental or behavioural action stemming from that schema. We therefore need a mechanism whose function is to support action-selection which responds simultaneously to the dictates of two separately coded schemas, and

which finds a compromise between the two. This will be called *dialogical processing*, and with the benefit of this mechanism, a schema-based cognitive architecture is enabled, more or less effectively, to respond to context while leaving its schemas intact. Where this mechanism is not active, and action-selection considers only one schema at a time, this will be called *monological processing*. (These terms were first used by Bakhtin 1934–1941/1981, but I shall not attempt to relate the two usages here.)

We now turn to some structural points concerning situations in which the accommodation of context modulates the selection of mental or behavioural action. One set of factors, which will be called a *matrix*, may give reason by default for the choice of one action, while additional consideration of a further set of factors, a *context*, may give reason to choose an alternative action. We thus have a structure of four elements, $<M, A_1, C, A_2>$, in which selection of the *default-action* (A_1), on the basis of the matrix (M), is defeasible and may be overridden by additional consideration of the context (C), leading to selection of the *alternative action* (A_2).[1]

In a schema-based architecture, then, the handling of such structures will depend on how they are coded and processed. If the matrix and context are coded together in one schema, then the context may successfully be accommodated through monological processing. If, however, matrix and context are coded across two separate schemas, then dialogical processing will be required in order to engage and accommodate the context during processing of the matrix.

This model,[2] sketched briefly above, was developed in response to the pattern of similarities and differences obtaining between typical and autistic cognition and behaviour. The syndrome of autism was identified by Kanner (1943) in the United States and by Asperger (1944/1991) in Vienna, and involves peculiarities in flexibility, spontaneity, socialisation and communication, together with the occasional appearance of special abilities.[3] In particular, the syndrome involves 'non-mentalising' deficits which do not essentially involve other people (such as general rigidity in behaviour and resistance to change); and 'mentalising' deficits which do (such as impaired performance in 'mindreading' in which one person works out the 'mental states' of another: what they believe, feel, want, etc.) My contentions are (i) that in typical cognition, dialogical processing supports successful performance in both the non-mentalising and mentalising tasks which present challenge in autism; and (ii) that in autism there is a *monological bias* in processing style which produces a large part of the observed range of deficit. One aspect of this model is that it implies that any given task which may be addressed dialogically may in princi-

ple also be addressed monologically (since the relevant matrix and context may be coded together into a single compound schema). Thus the bias in processing style attributed to autism does not necessitate deficit, and may be compensated through the development of specialised, compound schemas containing extra detail and exception-clauses. With regard to mental simulation, as explained below, the present model identifies this as one of two routes to mindreading, and situates simulation within a larger range of aptitudes which in typical cognition is supported by dialogical processing. The issue of mental simulation can of course be considered without any reference to the syndrome of autism, but this is to deny ourselves the negative illumination provided by functional impairment.

2. General cases

2.1 Rain Man at stop sign

We start with an illustrative example of autistic cognition, from the film *Rain Man*. The character with autism is crossing a busy road at a stop-sign. Half way across, the sign changes to 'Don't Walk', and he stops and stands still, despite the fact that he is blocking the road and the drivers start hooting. This situation involves a structure in which the *matrix* is the rule that the sign should be obeyed, and the *context* is the fact that the character in the story is standing in the middle of a busy road. Consideration of the matrix alone produces the *default action* of stopping walking, while additional consideration of the context will produce the *alternative action* of moving to one or other side of the road. Success through monological processing requires that the matrix and context be coded together in one schema, producing a rule something like 'Obey the sign, unless you are already on the road, in which case continue walking'. A dialogical strategy, on the other hand, achieves success by balancing the demands of matrix and context, without requiring that they be combined into a single rule. (An evident problem with the monological strategy here is the difficulty of producing a rule which is elaborate enough for all eventualities. The rule above, for example, does not mention whether you are able to move quickly, how far across the road you are, whether the traffic looks very eager, and so on.) The point for autism, then, is that a cognitive system with a monological bias in processing style will be reliant on the existence of a single, internally adequate rule, and in the absence of such a rule will manifest context-insensitivity of the type seen in this example. And the point for typical cognition is that, al-

though the monological strategy is available in principle, dialogical processing can be used in order to 'bend' the selection of action stemming from whatever schemas are available, thus producing adaptation to a situation 'on the fly' without reliance on a specially elaborated knowledge-structure.

2.2 Westminster Abbey

There now follows an anecdotal case.

> A schoolgirl with autism on an outing was found drawing graffiti on the walls of the Houses of Parliament. The teacher told her that she must not do this. But soon afterwards, when the party visited Westminster Abbey, she was found drawing graffiti on the walls of this building too, and was surprised when told that she had already been asked not to do this.

What was needed here was active search for a *context* (the fact that public buildings should not be defaced) to compliment the *matrix* provided by the teacher's instruction. This would have placed the teacher's instruction in a fourfold structure as described above, and thus when encountering the walls of the next public building the *default action* of drawing would have been overridden by the *alternative* choice of refraining from doing this. A successful monological response to the situation at Westminster Abbey would require the existence of an adequate rule saying something like 'Walls can be drawn on, but not if they belong to public buildings'. But this rule was not stated by the teacher, and in order to formulate it, active dialogical processing was required in order to connect context and matrix in the situation at the Houses of Parliament. Of course, ignorance of the contextual fact that public buildings should not be defaced would produce the same result as monological bias (in which the fact is known but ignored), but the occurrence in autism of patterns in which a person is aware of a contextual fact but ignores it in action-selection are too common to allow the general application of this sort of explanation. The point, then, for typical cognition is that the generalisation of rules in learning may require active dialogical processing in which a given matrix is connected with a relevant context.

2.3 Ignoring policemen

In another case, an able woman with autism had been told by her mother not to talk to strange men in the street. On one occasion, she was approached by a group of policemen who were concerned for her welfare, and following her

mother's rule, she simply ignored them. As a result, the policemen were annoyed and arrested her. As she herself said later, what she had needed was a rule such as 'Don't talk to strange men in the street, unless they happen to be policemen'. In the absence of such a rule, what was required was that action stemming from the *matrix* of the available rule be modulated in response to the *context* given by the fact that the strange men at issue were policemen. That is, it was required that a single schema should not be given absolute cognitive authority in the selection of action, as happens in strongly monological processing, and that the dictates of the relevant schema be moderated in response to context.

3. Mentalising cases

The three situations described above all involve other people: drivers, a teacher, and a group of policemen. It might be argued therefore that this (rather than the fourfold defeasibility structure identified above) is the factor which creates difficulty in autism. However, a range of experimental evidence indicates that individuals with autism show context-insensitivity in situations of this structure whether or not other people are involved. Thus, Frith and Snowling (1983) report context-insensitive interpretation of single words in sentences; Turner (1999) reports inability to connect an object such as a newspaper with the contexts of alternative, non-obvious uses to which the object might be put; and in a range of 'executive' tasks, individuals with autism show fixation on a rule or procedure, producing prepotency and perseveration in relation to an additional requirement of the task (see Russell 1997). This suggests that the context-insensitivity manifested in autism is not confined to the mentalising cases in which we need to work out the mental states of other persons. The question, then, if our model is to bridge over to these mentalising cases, is how a capacity for dialogical processing can support mindreading, and here we focus on an examination of the structure of the now famous 'false belief tasks'.

3.1 The false belief tasks

Wimmer and Perner (1983) describe an experiment in which the participant sees an object being moved from one location to another while a 'protagonist', Maxi, is absent, and the participant is then asked to say where the protagonist thinks the object is (or where he will look for it). The correct answer, then, is 'In location-1', and the answer 'In location-2' fails to accommodate the perspective

of the protagonist (who is under-informed about the situation). These authors found that success in this task emerges around the age of four years in typically developing children, as has been widely replicated.

Peterson (1997) and Peterson and Riggs (1999) propose that there are two alternative strategies for success in this task. In the first strategy, we reason from a principle of folk psychology such as 'If a person is absent when an object is moved, and does not have reason to expect the change, then he or she will assume that the object is still where it used to be'. In the second strategy, we operationalise Maxi's epistemic perspective by imposing the instruction 'Ignore the change in location' during the process of answering the base-question 'Where is the object?' In this way we modulate the behaviour of our own cognitive system so as to produce a virtual and temporary model of the behaviour of another cognitive system, and we give the answer 'In location-1', thus responding from Maxi's perspective rather than from our own. This gives needed detail to the notion of mental simulation as proposed by Gordon (1986) and Heal (1986), and articulates a strategy whereby mindreading in the false belief tasks can be achieved through adaptive modelling of another cognitive system. Peterson and Riggs (1999) propose that this is the strategy which children typically try to use when presented with this false belief task, and that the younger children fail because of the conflict-handling required to sustain the imposed instruction (since it collides with the fact that the object has been moved, as is represented in the child's knowledge-base). And support is given to this proposal by the finding, reported by Riggs, Peterson, Robinson, and Mitchell (1998, experiments 1 and 2), that in typically developing children there is a strong correlation between performance on mentalising test-questions such as 'Where does Maxi *think* the object is?' and on matched, non-mentalising test-questions such as 'If Mummy had not moved the object, where would it be?'.

On the present model, the strategy of reasoning from a principle of folk psychology is monological, since it relies on a single adequate rule (as may indeed be useful in autism and non-autism alike). The second strategy is dialogical, in that a *context* (the instruction 'Ignore the change in location') is engaged during the processing of a *matrix* (the set of relevant facts known to the child), while answering the base-question 'Where is the object?'. We thus expect deficit in performance on this task in autism, as is reported by Baron-Cohen, Leslie, and Frith (1985) who report that in autism the usual developmental shift does not take place, and children above the (mental) age of four years continue to answer according to where they know the object really to be rather than according to the perspective of the protagonist.

The case of autism, however, is not simple, since some children do pass this task, and as reported by Bowler (1992), high functioning adults with a diagnosis of autism may pass second-order false belief tasks. Frith, Morton, and Leslie (1991) argue that such success is achieved by 'hacking out' a task-specific solution, and Bowler suggests that it is achieved through an alternative route which may manifest differences in timing when compared to typical performance. On the present model, such success in autism is explained as occurring through recourse to the monological strategy of reasoning from a principle of folk psychology, thus achieving mindreading without reliance on dialogical processing. And where this route is taken, our model predicts a lack of the flexible, cooperative cognition supported by dialogical processing, but does not rule out the possibility of interest in other persons and the ability to predict their behaviour on the basis of learned principles.

This raises the question of how the relevant principles are acquired in the first place. If basic understanding of the nature of a personal perspective is missing, then it seems likely that the relevant schemas will not develop. Vygotsky (1978) and Hobson (1993) have argued that the development of understanding of alternative perspectives is achieved through early interactive experience of shared, intersubjective thought and feeling. We here make the *developmental conjecture* that such early experience is reliant on dialogical processing, as defined above. And in this case, our model predicts that a strong monological bias will produce a blocking effect whereby basic understanding of alternative personal perspectives does not develop, and hence the formulation of the relevant principles of folk psychology will be blocked also. This suggests a split in performance within the syndrome of autism, whereby (i) a moderate monological bias will produce general insensitivity to context, together with reliance on scripted rules for some situations, and (ii) a strong monological bias will produce the same general insensitivity to context, together with a 'blindness' to an area, such as the social domain, in which the acquisition of basic understanding (which might otherwise be converted into scripted principles) is blocked. The point is speculative, since our developmental conjecture of the priority of dialogical processing in the development of basic understanding of issues of personal perspective is in need of proof; but it is consistent with the observed difference within the autistic spectrum between individuals who are socially active but 'odd', and those who are 'cut off' and 'oblivious to other people'. And hence, if correct, this conjecture may have application to the observed differences between the versions of the syndrome of autism identified by Asperger (1944/1991) and Kanner (1943) respectively.

In any case, in developing a theory of mindreading, the phenomena of autism force us to engage with the differences and splits in performance which are found within this syndrome. In particular, while many individuals with autism fail false belief tasks, some pass, and there is reason to think that those who pass do so through a specialised strategy which lacks flexibility and generalisation. Thus, for example, Happé (1994a) reports a study in which it was found that participants with autism who passed a second-order false belief task made errors in another task involving people's mental states. This point is consistent with the observation that in many areas people with autism do not show a blanket deficit, but rather manifest partial aptitude which differs in character from the typical case. Thus, social skills may be developed, but these tend to have a brittle, scripted quality, showing non-adaptation to unexpected context; and pretend play may be manifested, but this generally happens through copying or in response to an instruction (see Lewis & Boucher 1988, 1995), and lacks spontaneous extension and flexibility. On the present model, the individual with autism is confined to the monological style of processing, with its reliance on single, internally adequate schemas for particular tasks. And while this strategy is viable in principle, and may produce asset were the accommodation of context is detrimental, the inherently limited nature of any individual schema will produce context-insensitivity of predictable sorts.

4. Conclusion

In addressing the issue of mindreading, we need to consider the phenomena of deficit found in the syndrome of autism. These phenomena are differentiated, and involve the appearance of aptitude of a particular character alongside more general deficit. This calls for an account which acknowledges at least two alternative routes to mindreading, and which addresses the particular nature of aptitude where this appears in autism. The present model posits a schema-based cognitive architecture supporting both monological and dialogical processing styles in the selection of mental and behavioural action. In case of monological bias, then, an individual will be more or less confined to the monological route, which in the case of mindreading consists in 'theory-based' reasoning from principles of folk psychology. On this model, mindreading through mental simulation is identified as relying on dialogical processing, and is analysed as requiring dual response to a context generated for the other person together with the matrix of our own general knowledge. In this way, we operationalise another person's perspective by temporarily causing our own selec-

tion of mental action to mimic that of the other person. This model correctly predicts a correlation found between performance on 'mentalising' and 'non-mentalising' test-questions presented to typically developing children, as discussed above, and is consistent with the pattern of interwoven deficit and aptitude found in mindreading in autism, and with the fact that in autism deficit in mindreading appears alongside a general problem in the accommodation of context. With regard to the place of simulative mindreading in mental life, the present model locates this within a larger picture: mental simulation relies on dialogical processing, which also supports a range of context-guided adaptive behaviour, and rather than being the only route to mindreading, this relatively flexible strategy is partnered by the more rigid monological strategy of inference from known principles of folk psychology. This has implications for the relation between typical and autistic cognition, since it presents the autistic case, not as a deviation from the typical case, but as a part of it, since monological processing consists in the unmodulated activity of the basic function of action-selection in a schema-based architecture; and conversely, when we employ mental simulation, we add to rather than abandon the route to mindreading which is supported by this basic function. If, however, our developmental conjecture is correct, then these alternatives are not disjoint, and an early and especially strong bias towards the monological style of processing may in turn block the possibility of its own efficacy in mindreading.

Notes

1. The notion of defeasibility appears in many fields. In his discussion of legal reasoning, Hart (1949) took the term from the law of property in which a right over property may be assumed to exist, but may be overridden when it is discovered that another party has a conflicting right; attempts have been made in artificial intelligence to accommodate defeasibility (see Minsky 1975 and Pollock 1987); and in formal logic, inference which is defeasible is characterised as non-monotonic (see Hodges 1977).

2. For helpful comments during the development of this model, my thanks go to Oliver Black, Antoni Diller and Rita Jordan. My thanks also to the pupils and staff of the Language Unit of the Parkview Clinic Birmingham.

3. See Happé (1994), Frith (1989) and Rutter (1999); World Health Organisation (1990) and American Psychiatric Association (1994).

References

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.: DSM-IV). Washington, DC: American Psychiatric Association.

Arbib, M. A. (1989). *The Metaphorical Brain 2: Neural Networks and Beyond*. New York: Wiley-Interscience.

Asperger, H. (1944/1991). 'Autistic psychopathy' in childhood. In U. Frith (Ed.), *Autism and the Asperger Syndrome* (U. Frith, Trans.). Cambridge: CUP.

Bakhtin, M. M. (1934–1941/1981). *The Dialogic Imagination: Four Essays by M. M. Bakhtin* (C. Emerson and M. Holquist, Trans.). Austin: University of Texas Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition*, *21*, 37–46.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Bowler, D. M. (1992). 'Theory of Mind' in Asperger's Syndrome. *Journal of Child Psychology and Psychiatry*, *33*, 877–893.

Frith, U. (1989). *Autism: explaining the enigma*. Oxford: Basil Blackwell.

Frith, U. & Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *Journal of Developmental Psychology, 1*, 329–342.

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder: autism. *Trends in Neuroscience*, *14*, 433–438.

Gordon, R. (1986). Folk Psychology as Simulation. *Mind and Language, 1* (1) 158–171.

Happé, F. G. E. (1994). *Autism: an introduction to psychological theory*. London: UCL Press.

Happé, F. G. E. (1994a). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*, 129–154.

Hart, H. L. A. (1949). The Ascription of Responsibility and Rights. *Proceedings of the Aristotelian Society*, *49*, 171–194.

Heal, J. (1986): Replication and Functionalism. In J. Butterfield (Ed.), *Language, Mind and Logic*. Cambridge: Cambridge University Press.

Hobson, R. P. (1993). *Autism and the Development of Mind*. Hove: Lawrence Erlbaum Associates.

Hodges, W. (1977). *Logic*. Penguin.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child, 2*, 217–250.

Korf, R. E. (1988). Search in AI: A Survey of Recent Results. In H. E. Schrobe (Ed.), *Exploring Artificial Intelligence*. San Mateo: Morgan-Kaufmann.

Lewis, V. & Boucher, J. (1988). Spontaneous, instructed and elicited play in relatively able autistic children. *British Journal of Developmental Psychology, 6*, 325–339.

Lewis, V. & Boucher, J. (1995). Generativity in the play of young people with autism. *Journal of Autism and Developmental Disorders*, *25*, 105–121.

Minsky, M. (1975). A Framework for Representing Knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill.

Peterson, D. M. (1997). Representação, Inferência e Simulação em Tarefas Envolvendo Crenças Falsas. In M. E. Q. Golzales, C. A. Lungarzo, C. B. Milidoni, A. Pereira, & M. B. Wrigley (Eds.), *Encontro com as Ciências Cognitivas*. Marília: Universidade Estadual Paulista.

Peterson, D. M. & Riggs, K. J. (1999). Adaptive Modelling and Mindreading. *Mind and Language*, *14* (1), 80–112.

Piaget, J. (1937/1955). *The Child's Construction of Reality* (M. Cook, Trans.). London: Routledge and Kegan Paul.

Pollock, J. L. (1987). Defeasible Reasoning. *Cognitive Science*, *11*, 481–518.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, *13*, 73–90.

Russell, J. (Ed.). (1997). *Autism as an Executive Disorder*. Oxford: Oxford University Press.

Rutter, M. (1999). The Emanuel Miller Memorial Lecture 1998, Autism: Two-way Interplay between Research and Clinical Work. *Journal of Child Psychology and Psychiatry*, *40* (2), 169–188.

Shank, R. C. & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Hillesdale, NJ: Erlbaum.

Turner, M. A. (1999). Generating Novel Ideas: Fluency Performance in High-functioning and Learning Disabled Individuals with Autism. *Journal of Child Psychology and Psychiatry*, *40* (2), 189–201.

Vygotsky, L. S. (1978). Internalisation of higher psychological functions. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in Society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

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.

World Health Organisation. (1990). *International Classification of Diseases* (10th revision: ICD-10). Geneva: World Health Organisation.

Reply to Donald M. Peterson

Pierre Livet
University of Provence, Aix-en-Provence

Peterson presents a very interesting explanation for why children of less than four years have deficits in the false belief task and other similar tasks: they have difficulties with the revision process that such tasks require. I fully agree with this hypothesis. But its content has to be specified in two ways.

First, is the Theory Theory approach necessarily a "monological" one, and the simulation approach a "dialogical" one? The answer to this question is not obvious. Our theory about the mental states of other people may be a default theory ("normally, people in such a situation believe B, but if some other context is added, they believe not B"). On the contrary, in order to simulate other people, we have to know how to behave in the relevant situation, and we could not know how to do this, if we had not previously been in a similar situation. Besides this, in order to block our program of action when the context activates the "additional consideration" that overrides default-action A_1 and triggers action A_2, we must not only have learned, by experience, action program A1 and action program A2, but must also have had experience of the revision program that triggers the shift from A1 to A2. This is more demanding than the Theory Theory script. Moreover, if simulation implies motor imagery, it is subject to motor constraints, and is less flexible than the more general and unspecified belief theory. Of course, this difference has implications for experimental observations and tests of the two theories.

Second, revision is not a simple matter. What is really at stake here? Is it revision, or updating? According to Peterson, in the false belief task, children have to cancel one of their cognitive routines, or one of the premises they use for the inferences drawn "on line", while observing Mummy baking the cake, Maxi and the smarties, or similar situations. For example, Mummy takes the chocolate used for the cake and moves it to the cupboard. We then ask the chil-

dren the counterfactual question: "Suppose Mummy had not baked the cake, where would the chocolate be now?". According to Peterson, the children are supposed to revise their premises by ignoring the premise that the chocolate was moved to the cupboard. Maybe things are not so simple.

Let us admit for the sake of the argument the Stalnaker-Lewis framework of conditionals, and the classical relation between counterfactuals and revision. In order to evaluate whether a conditional counterfactual is true, we imagine a set of possible worlds in which the antecedent of the counterfactual is true. Then we select in this set the worlds closest to the world of reference – usually, the actual world – and check whether the consequent is true in this subset. On the other side, revision can be considered as the operation of adding to our belief basis a new information, detecting some inconsistency, identifying the sets of premises – that is, the different combinations of premises – that could be responsible for this inconsistency. Finally we choose to cancel the smallest set of premises. If there are several minimal sets, we assign to each premise an epistemic priority, and cancel the set of premises that has the lowest priority.

One can see that revision and counterfactual reasoning imply the definition of a distance (the "worlds closest to the actual one") that can be an order ("the lowest priority"). Complications arise however when the revision is in fact an *updating*. In such a case, we cannot assume that the world has not changed, and that the source of revision is only a change in our information state. We have rather to consider the possibility that the new information results from a change in the world. Moreover, we also need to consider different sets of possible worlds in parallel: one in which the world did not change but where our informational state did, and another in which the world changed. When updating, we cannot get a unique solution. We have rather to take into account two or several different hypothetical solutions and to reason in parallel along these various different lines. Unsurprisingly, psychologists have shown that we are very bad in this kind of disjunctive reasoning.

When a real scene is presented to the children, like in Maxi and the smarties tasks, updating seems to be the rational strategy of reasoning when the children are asked what will be Maxi's behavior when he comes back in the room and looks for his chocolates. If Maxi is a rational puppet, he cannot exclude that something has changed in the room during the time he could not see what was happening in it. So he has to consider in parallel the possible worlds where nothing has been changed, and those where something has been changed. His rational belief has a disjunctive form: either no change happened and there are smarties in the box, or something has changed and I do not know what is in the box.

When the children are told the story of Mummy baking the cake, things seem a bit different, because an implicit rule of communication requires that if some relevant change has happened, the person who tells the story has to mention it (this rule is regularly violated in detective novels). So the rational strategy involves revision rather than updating. But notice that, in this experiment, children are asked where the chocolate would be if Mummy had not baked the cake. Rationally, if Mummy had not baked the cake, some change would have occurred in the world of the story. According to the Gricean presupposition, nothing is supposed to have changed except the fact that Mummy has not baked the cake. But "baking the cake" is the whole story! So if Mummy has not baked the cake, nothing is presupposed about what she could have done instead. In particular, Mummy could have moved the chocolate to the cupboard while performing another cooking task. The rational children are again required to adopt the updating strategy, and their answer again has to be disjunctive: if Mummy had not baked the cake, either she would have not moved the chocolate to the cupboard or she would have moved it for other reasons.

So less than four-year-old children fail to achieve not so much revision as updating, which apparently psychologists fail to do as well. An heterodox hypothesis could be then that children have difficulties not in achieving a simple revision, but in ignoring the difficulties of the updating task, and in reducing it to the relevant kind of revision. They have not yet learned that the socially acceptable way of dealing with this kind of problem consists in collapsing updating problems onto revision ones. Faced with this difficult problem, they stick to their own unrevisable information: the smarties have been changed, the puppet has been moved from one box to another, and so on. When determining what are the subjects' other beliefs, their strategy might be to rely on the facts of the actual world.

Of course this is a provocative hypothesis, and I have no intention to defend it in this form. However some parts of it are true, I think: adults have acquired a competence in simplifying situations, in considering them as typical stories (reducing in this way updating to revision); children have another and even more drastic way of simplifying the same problems. If this is right, what are the consequences for the debate between the simulation theory and the Theory Theory?

At a first glance, it seems that it vindicates the Theory Theory against the simulation theory. The competence for simplifying situations and consider them as typical stories requires the mastery of a theory about situations, the theory that treats them as stories. Adults possess this theory and children do not. But at a second examination, simulation theory and Theory Theory could

be both seen as compatible with the contrast sketched above. First, as children do not have the ability to reduce situations to typical stories on their own, they cannot imagine being in the shoes of other people and simplify situations in this way. Second, how can adults come up with such a simplification? Maybe not because they have a theory that tells them that situations are reducible to stories. Such a theory would be dangerous. But because they can simulate the reasoning of people engaged in a story, not in a real situation. When faced with a situation that would require a difficult updating, and if this task does not require real and important practical decisions, if the task can be done "for fun", in isolation from the practices of every day life, they simulate a revision task instead of an updating one. This allows them to give a definite answer instead of an uncertain and disjunctive one. Here the role for theory boils down to identifying the relevant kind of contexts isolated from ordinary life, and all the remaining work is done by simulation.

Surprisingly enough, such a kind of behaviour is very similar to "pretending" in young children. Adults play the game of revision instead of real updating, as children play their pretence activities instead of real ones. But the pretence activities of children are much less constrained than adults' ones. Children can change the rules of the game at once, ignoring the requirement of introducing explicit revision of the rules, as adults do. Maybe children fail to conjoin their pretence competence with their ability for dealing with real things, while adults build an intermediate world by reducing the real problem of updating to the real but simplified problem of revising.

Notice that, as autistic children are notoriously bad in pretend-play (but rather good at detecting real details), they cannot use their pretence capacity to build such an intermediate world. This is why they fail the false belief task, except when it is transformed in a memory task about different photos supposed to represent the same scene at different moments and from different perspectives.

Chapter 8

Can "radical" simulation theories explain psychological concept acquisition?

Joëlle Proust
Institut Jean-Nicod, CNRS, Paris

The question of understanding the behaviour of other agents in a psychological or mental way divides into two sub-questions:

1. How can a human child above 3 and a half or an adult attribute certain *kinds of psychological states* rather than others to people, i.e. a state of belief, rather than of desire, hope, fear, etc.?
2. How can such an interpreter find out correctly what is the *content* of the others' psychological states, i.e. a belief that P, a desire that Q, etc.?

Both Theory Theory (TT) and simulation theory (ST) aim at answering these questions, but each actually only offers one part of the explanation. TT explains the capacity of using psychological concepts as a result of an innate, specialised module or as a result of a general process of theory building. But it does not explain how psychological concepts may adequately be applied to specific contents.[1] ST explains how an interpreter can grasp the specific content of the other's psychological state, while leaving in the dark how a simulator can understand the various psychological attitudes that other subjects entertain towards that content.

In this paper, I will concentrate on the response offered to the latter problem by Robert Gordon, who espouses a simulationist view on how understanding others is achieved. This author indeed pushes to its limits the possibilities of *radical simulationism,* i.e. the view that mentalisation is a process requiring initially no possession of psychological concepts. It is interesting to see exactly what this theory can accomplish, and if it finally fails to explain mentalisation, *why* it does.

Radical simulationism: Main claims

The basic claims of radical simulationism will be better articulated against the background of a weaker version of ST, such as the one defended by Alvin Goldman (Goldman 1989, 1992, 2001). According to Goldman, simulating another agent consists in taking her point of view on the world, and in producing "off-line" the various responses (decisions, emotional states) that one would produce in such a situation if it was actually present. The basic feature of simulation, considered as a representational process, is that it is *isomorphic* to the target-process, which it is meant to mimic (Goldman 1992: 108).

According to Goldman, simulation is thus a three-step process. (1) Projecting into someone else's situation. (2) Appreciating introspectively what it is like to be in that situation. (3) Applying whatever psychological concepts available to categorise the feelings, desires and beliefs and other evaluations made in the second step, and attribute them to the simulatee. Simulating, in such a theory, requires a prior expertise in mental state attribution. You must be able to recognise, say, a specific emotion or a belief in yourself to be able to ascribe it to another person. You must have categorised your own psychological states to recognize them as applying to other agents.

Robert Gordon diverges from Alvin Goldman's view on simulation on three major counts.

1. The first point of divergence has to do with the kind of *projection* needed to simulate others. Is the target of a projection a mind, a situation, an idiom? It will be useful here to contrast Gordon's view on projection with Quine's and Goldman's. Projecting oneself, in Quine's approach, implies that the simulator should imagine what it is like to be in another's state of mind. By this Quine understands that the simulator reconstruct the *system of sentences held true* by the simulatee. This process consists in translating the other's idiom into the simulator's, and in deriving the sentences that hold in the idiom modified in the appropriate way, according to a rationality principle:

> We project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then we say what, in our language, is natural and relevant for us in the state thus feigned.
>
> (Quine 1960: 92)

Alvin Goldman would agree that projection has to do with imagining being in another's *state of mind*, but he would deny that projecting always involves an exercise in translation. He explicitly allows simulation to account for non-

propositional mental states like pains and itches (1995: 84). Goldman maintains that simulating requires *introspecting* what oneself would feel and believe if one was in such a situation, and ascribing similar contents to the simulatee.

Robert Gordon does not see projection as allowing any kind of introspection, nor would he indeed accept without qualification the view that in simulation, an agent generally projects into somebody else's *state of mind.* On the other hand, Gordon does not take Quine's view on projection as a translation from an idiom to another. Projecting consists rather in recentering oneself into another's location and concerns through what Gordon calls an "egocentric shift". A "personal transformation" into the simulatee occurs, rather than a transfer into somebody else's mental states (Gordon 1995: 56). Simulation occurs through an imaginative transformation into the other, which allows the simulator to see the situation that the simulatee faces as *she* faces it. One might here object that it is not quite clear how such a shift may occur if not through some mental transfer. A plausible reconstruction might be the following. A simulator projects herself into a situation-type, i.e. in the situation that typically causes the mental state of the simulatee and makes it the type of mental state it is, rather than into the mental state itself (although projecting into a situation-type leads *de facto* to having some consonant set of beliefs and desires).

2. The second major point of divergence has to do with *mastery of psychological concepts being or not a prerequisite* for simulating others, and more generally, with the relationship between simulating and believing. In Alvin Goldman's view, a simulator needs to be able to identify conceptually the kind of the mental state that she introspects in the pretend mode, in order to assign it to the simulated person. Simulating presupposes further an ability to frame the proper kind of situation to be simulated, an ability that involves a considerable amount of knowledge of what others might believe and want about the situation.

Gordon recognises that simulating situations different from one's own sometimes requires a mastery of psychological concepts. He maintains however that there is a large class of simple cases in which no such mastery is required because no motivational or epistemic adjustment need to be done; furthermore, the capacity to simulate in those simpler cases might explain how psychological concepts are finally acquired. Early forms of simulation thus seem to him to require no mastery of psychological concepts, but rather, combined with "ascent routines", to allow a first crude way of categorising intentional states. The core idea of an ascent routine is the following. Phrases such as "I

believe", "I want", etc. are first uttered purely expressively, in the absence of psychological concept possession: beliefs, wants, desires are expressed through a linguistic form borrowed from ordinary adult conversation, without the child having any particular concept of what the corresponding propositional attitude is.

Ascent routine was first described by Gareth Evans, in his *Varieties of Reference*, (1982) in a chapter devoted to *self-identification*. "In making a self-ascription of belief, he writes, one's eyes are, so to speak, or occasionally literally directed outward – upon the world" (p. 225). To know whether I believe P, "I have to put into operation whatever procedure I have for answering the question whether P". There is a semantic ascent in this process, because from P, i.e. a fact which I take as holding in the world, I can derive a fact about myself, which I assert through the words "I believe that P".

Now as both Gareth Evans and Robert Gordon insist, using this procedure does not amount to a "full understanding of the content of the judgement" "I believe that P". For Evans, a subject who performs the ascent routine may still lack the mastery of the generality condition, which applies to psychological concepts such as "x believes that P". In other words, she may fail to understand that the content she believes, namely P, can also be believed by other people. She may also be unable to grasp the fact that a belief is a claim for truth, which can be supported or not by the right kind of evidence. What the ascent routine does achieve, is a way of bypassing the notion that information about the self has an inner source, to be accessed through introspection (Evans 1982: 230).

For ascent routine to work in the case of primitive simulations, one needs only accept that a subject who is able to "label" a belief or a desire in a real situation is also able to label those attitudes in pretence. Robert Gordon acknowledges that this early kind of simulation does not deliver a genuine, comprehending ascription of mental states, but he nevertheless claims that such simulatory practice may foster "greater conceptual understanding" (1996: 17). He suggests further that a process of embedding ascent routines in simulations "gives sense to the idea of a mental location". When simulating Peter, John projects himself in Peter's situation, and evaluates it from Peter's point of view. Projection amounts to a process of accepting propositions believed by Peter given a situation S. John thus gets a notion of "something's being a fact to Peter". Projection is normally followed by a deduction of those consequences of the accepted propositions relevant to Peter in S, and by an ascription to Peter of a belief in the propositions thus derived. By embedding ascent routines in simulation, Gordon means to refer to the iteration of the question "is it a fact that P?" in various perspectival situations. The simulator is able to store the vari-

ous responses to the same question using various individual-based situations: is it a fact that P for Peter? Is it a fact that P for Mary? Etc. What John finally learns is that some facts are acknowledged by most individuals, some others by few. What he will be unable to learn in this way, Gordon notices, is that there are facts *without* a mental location, i.e. facts considered by nobody, and about which the question of truth or falsity as a matter of fact does not arise.[2]

3. Reliability. The question that is raised by the role of practical simulation in mentalisation is of what makes self- and other- ascription of beliefs and desires generally *reliable*. A process is reliable if it tends to produce true beliefs. As Gordon observes (Gordon 1993: 45), reliability in self-ascription of mental states may be taken to stem from a decision procedure – provided by appropriate concept application in Theory Theory, and by introspection and categorisation in Goldman's approach. Gordon favours a view in which reliability does not rely on a decision procedure, but rather on a non-inferential, non-conceptual process: verbal labelling as described in the ascent routine provides reliability of self-ascription without any concept being actually called for. In his words, "reliability does not need theoretical knowledge".

Objections to radical simulation theory

A first well-known problem with the view that a simulator only needs to use the information contained in some perspectival situation is that some mastery of epistemic notions must be present for a subject to select, during deduction, those premises that are accepted by the simulatee. In other words, the notion of a situation relevant for belief ascription cannot be defined in purely extensional or subject-neutral terms. A simulator must have some notion of which premises will be used in a simulated piece of practical reasoning. For example, if we take the tube test,[3] which end will Peter choose? The unconnected opening just under the spot where a ball is launched, or the farther but connected opening? To predict adequately where Peter believes that the ball will go, I must include in my premises the proposition that Peter accepts the view that the ball has a causally restricted dynamics. If Peter accepts instead that a shorter distance between launching and picking up is a relevant element in the problem, my deduction will take another course. It is hard to see how such an appreciation of what premises are relevant can be effected without presupposing the very notion to be explained, i.e. the notion that a subject *believes* those propositions which he accepts and acts upon.

A second difficulty concerns the view on psychological concept acquisition taken to result from a diversified simulatory practice. Gordon takes "embedding" of simulations to consist in a rather straightforward inquiry of whether the fact P is holding from the various perspectives of individual thinkers. The simple structure of "embedding" which Gordon uses, however, falls short of providing a notion rich enough to prepare the understanding of a *mental* location. In his view, decentering one's perspective on a situation only requires changing places, in the literal sense of varying the spatial distance and perceptual access to objects, and keeping track of the various views on enduring objects and states of affairs. Spatial memory allows storing various points of view on one and the same situation, along with the specific properties associated with each one of them. Given an observer endowed with objectivity (i.e. capable of grasping the fact that perceived objects are stable individuals existing even when they are not perceived), a subject is in this analysis able to grasp perspectivalness. When I was there, the object looked round. Now I am here, and the object looks oval.

Let us grant that simulating a situation does not involve taking the point of view of any specific subject, but only presuppose access to the informational content which one gets from occupying various possible locations. It may seem that a simple spatial variation between what is seen from here, from there, etc., offers the required contrast between situational properties, and prepares the acquisition of the concept of belief.

But this seeming is illusory. It is a product of the silent use of psychological concepts in the description of spatial changes. The verb "seeing" carries the burden of conveying a psychological information on top of the information collected about the scene. More strictly described, spatial change appears no longer as sufficient to provide the sense of a mental location. Clearly, as Strawson (1959) showed, changing one's view on the world constitutes a necessary condition for representing stable entities that may exist unperceived. But Gordon's claim is that perspectival simulation is a sufficient condition for understanding mental perspective. Data from primatology support rather the idea that there is a cognitive gap between understanding perspective and understanding belief.[4] Non-human animals are quite good at predicting, in a simulatory way, which rival is likely to grasp some food, or to reach some prey, without having the least disposition to attribute to others the notion of "P being *a fact for X*" (or *X* believing P). Chimpanzees, for example, are certainly acting as if they had such a notion. But careful experimental work by primatologists (Povinelli 1996) shows that it is not the case. Although they can use some psychological information in the sense that they can follow the gaze of

conspecifics, and use that information to detect for themselves what there is to be seen, they are unable to understand that the other acquired some perceptual knowledge. Non human primates are unable to tell apart an informed partner from a naive one; they cannot infer directly knowing from seeing. Although they may learn how to hide some behavioural signs of their inner states, they seem to do so on the basis of associations between behavioural cues. Not: "If I shout, he will know". But: "If I shout, he will bite". Autistic children seem also to have trouble to infer from a perceiving subject that this subject knows perceptually something (Baron-Cohen & Goodheart 1994). On the other hand, they have no trouble finding out what will be on the picture, when the objects in front of the camera are being changed, a task which plausibly involves simulating prior perspectives (Zaitchik 1990; Leslie & Thaiss 1992). Mastery of perspectival information does not amount to mastery of epistemic perspective.

These findings tend to show that *extracting* psychological information – simulation being one way of doing so – may guide behaviour without allowing the agent to understand that the information used is indeed mental. For all the animal knows, so to speak, the information used is behavioural. It has to do with the types of things others can do in a particular situation, not with the types of things they can believe and want to do. Therefore changing perspective through simulation does not offer *per se* any clue about psychological capacity.

Let us come finally to Gordon's view on reliability. The notion of reliability that Gordon has in mind does not amount to a capacity of providing *rational grounds* for what one says or does. It has to do with the fact that the agent, while being or simulating being in a situation, will *normally* identify the right content of his/her belief or desire. After some linguistic training, children are able to utter "I want a banana" when and only when they want a banana. Doing so does not amount to a "genuine self-ascription of desire", Gordon acknowledges, because they may still lack a concept of desire (as well as a concept of themselves). This analysis illustrates the fact, stated above, that simulation theory is better equipped for explaining how particular psychological contents can be understood than for categorising the attitude which the simulated subject has towards that content. True, a child manages to convey that she wants a banana, rather than an orange; doing so, she does not need to know that there is something she wants, rather than hopes, regrets, believes, etc. But this simply reflects a fact about training; a trained animal reliably does what it was trained for doing. Why should this tell us anything about psychological concept acquisition?

Therefore the reliability of the mechanism for *uttering* sentences that *reflect* mental states (but actually do not *refer* to them) should be contrasted with the reliability of a mechanism for self-attributing mental contents. Younger children may communicate what they want reliably, while utterly lacking a capacity for self-ascribing correctly psychological state types. Is the first kind of reliability a step towards acquiring a reliable mechanism of the second type? We shall struggle with this problem at the end of the chapter.

Simulation reconsidered

Gordon's radical simulationism stems from three important intuitions. The first is that simulation may be performed in a context-driven and agent-neutral way. The second is that simulation is a very basic, multi-purpose psychological process, that does not necessarily involve psychological concepts. The third is that simulation is harnessed to a practical reasoning system. From these three basic intuitions, the challenge consists in showing that simulation may indeed lead to psychological concept acquisition, or at least may provide an essential element in mentalisation.

A. Simulation as a context-driven process

Any kind of action involves representing a context, or a situation, in which the corresponding type of action normally develops. When an agent aims at producing a certain result by acting in a certain way, he/she has to represent what can be called "the canonical context" for that action. Such a canonical context includes a certain type of movement or active involvement, as well as the relevant objects and properties (or the most salient) for reaching the given goal. If an agent's goal is to get food, for example, she will have to represent the canonical context most appropriate for that goal, which involves both perceptual cues and propositional knowledge (including concepts and inferential relations between them): is this cake on the kitchen table available? Is it fully cooked? Should I go instead to the bakery? Etc. Memories of prior successful actions determine what constituents will be included in a given canonical context. A context may be represented, in a few cases, in an essentially nonconceptual way (for example, when it comes to a particular step to be executed in dancing with a given rhythm and tempo). It may also, in other cases, involve mostly conceptual knowledge (such as, for example, the kind of events to expect when taking an exam).

Most contexts involve both kinds of contents, conceptual as well as non-conceptual. Such an "action space"[5] organised in the perspective of an acting subject is the fundamental representational structure activated and exploited in simulation. As simulation theorists have insisted, a capacity to simulate necessarily depends on the experience acquired by a subject: a given representation of a context includes elements that may not play a role in practical reasoning, or may be supplemented, replaced by other items, etc. When planning an action, an agent will need to simulate various contexts for reaching her goal, and various courses of action within one context. In the simulative phase, the agent needs to articulate specific perceptual expectations with conceptual-instrumental reasoning. A young child may at first have trouble identifying the essential ingredients constituting a specific canonical context. Observing repeatedly the actions of other agents, listening to stories, engaging in pretend play, are the various means which help a child engage actively in context-typical simulations in ways that are more and more specific and appropriate.

Empirical evidence tends to show that human agents represent actions, motor sequences and instrumental relations between means and goals (i.e. canonical contexts) in a way that is "agent-focussed" and "subject-neutral". By these expressions is meant that an observed action is represented in the perspective of an agent, rather than in the perspective of the particular person who observes the action (who may be the acting self, but also somebody else). Granted that a canonical context of action articulates a goal, means-ends relations and specific perceptual cues, it seems particularly important that the specific orientation of the observer's body does not present a systematic obstacle to understanding the actions of others. Brain imaging studies (Decety et al. 1997) support the view that, in both cases, the dynamic representation of a context – its simulation – are performed under the agent's point of view. *Even when someone else's action is observed, the observer activates for himself the dynamical representation of the corresponding action in the neural structure that normally allows execution*.

This does not mean that the action is fully performed by the observer; nor that understanding another subject's action presupposes introspection of the same action performed by self; this only means that the "active" coding of an action is involved in performing as well as in observing it. A subject who has to identify whether a presented hand is a right or a left hand will move mentally her own hand – at a subpersonal level. A subject watching someone take an object with two fingers will activate a similar pattern in her *premotor* area. Seeing an action seems to constitutively involve a disposition to act in a similar way.[6] Although observing an action rarely evokes a conscious disposition

to act, there are circumstances – e.g. observing a soccer game or watching a thriller movie-, in which the observer clearly feels immersed in the action he watches, and actively projects into it in a way that is consciously accessible. He catches himself running and jumping.

B. Simulation as a psychological process

From a psychological point of view, simulating consists in running "off line" stored information. Such an information may be perceptual and either refer to facts retrieved in an imagistic (analogic) mode, or to particular aspects of those facts (one can simulate having an experience of blue, of heat, of pain, of drinking coffee etc.). It may be pragmatic or praxic,[7] with again the same contrast between whole contexts and individual properties: one can simulate oneself as doing this and that, or as moving in a particular way (in a more or less richly imagistic way). It can also be propositional, and simulation can be conducted on the basis of compositionality of senses – as is the case in planning. The feature common to these simulatory episodes is that their contents are represented as outcomes of a subject's concrete abilities or doings rather than in an abstractly inferential way. In this respect, simulation has much to do with episodic memory (Tulving 1983). For example, while simulating a future action (an activity that essentially involves concepts and propositions), the planning agent considers the consequences of a selected context in a practical way (how the situation will then appear to her), rather than as detached predications (made so to speak "from nowhere").

Simulating is thus a basic kind of dynamic memory process used in various areas of procedural knowledge, that cannot be divorced from the conceptual capacity used by a subject to categorise and infer external events and properties. The difference between semantic memory representation and simulation is simply that the former proceeds in a purely conceptual way, while the latter represents contexts in a mixed way. It uses both *concepts* to categorise a given context and *qualitative, first-person experience* to characterise the feelings and emotions relevant for the overall evaluation.

A second important point is that in each of its specific domains of application, simulation can also be used in many different ways according to what we might at least *prima facie* characterise as the reasoning capacity available to the subject. Through simulation, as we already saw, various kinds of experienced canonical contexts with their associated inferences are extracted from memory and combined. In some systems it will be mainly used to revive past embodied situations; in others, it will be used in reasoning about future or counterfac-

tual contexts.[8] As several theory-of-mind-theorists observed,[9] simulating can be of any help to a mentaliser only if it allows deploying the various kinds of inferences relevant in the simulated situation. The scope of reasoning in an individual thinker might here be constrained by the executive dispositions of the individual: those necessary for maintaining distinctive representations of one and the same context separate, and for exploiting the inferential relations across them.

C. Practical reasoning, counterfactual reasoning and simulation

There are two ways of simulating a counterfactual situation that need to be contrasted and explored further. Overlooking this difference may explain in part why Robert Gordon's radical simulationism fails to state the conditions that lead from simulation to mentalisation. To articulate this distinction, we need to explain what the *frame of reference* of a simulation is.[10] A frame of reference consists in the modality in which an agent deploys her representation of a given context. A frame of reference can be the real world (in that case the subject represents past or present, actually experienced context tokens or types). It can be a potential world: a future context token is represented as relevantly analogous to a past one. It can also be a non-real state of affairs, i.e. counterfactual or strongly counterfactual: in the latter case only, the subject represents explicitly the situation as part of a possible world distinct from the real world.

The capacity of mastering the various modal dimensions of thinking in distinctive frames of reference has interesting links with the capacity to *reason conceptually*. The generality principle, in virtue of which mastering a concept implies the ability to entertain "indefinitely many thoughts" in which this concept appears (Evans 1982: 105) would seem to require from any concept user that he should already be capable of counterfactual reasoning. This type of reasoning seems to be indeed a major way of exploring in thought the various combinations allowed by conceptual components. This requirement would have the unfortunate consequence to restrict concept use to those thinkers that are able to represent possible states of affairs in exotic worlds. For example, is it needed, to master the concept of a "unicorn", to understand sentences such as "had they been real, unicorns would have superseded horses in number"? A weaker interpretation seems compatible with the generality principle, a principle which, according to Evans himself, is "an ideal to which our actual system of thoughts only approximately conforms" (*ibid*. 105). In this weaker sense, mastering a concept would imply an ability to entertain thoughts involving that

concept inside at least one frame of reference. Accepting this view on concept mastery would lead to various extensions of "generality" according to the set of objects and contexts included in the world (or sets of worlds) represented in thought. This kind of distinction would account for the fact that a child may reason about unicorns inside the story which presents them as having some properties, while being unable to consider them "outside the story", as fictional entities with no direct causal impact on the real world.

What is true of concept use also applies to episodes of simulation. Not only because an episode of simulation often involves concepts, as we saw above. But also because simulation, being a species of representational thought, also possesses a frame of reference for its own sake, independently of the concepts it contains. It may be the case that imaging a shade of blue, for example, presupposes having actually seen the same shade (just as applying a concept to a new object presupposes having the concept). But nothing seems to restrict imaging to retrieving prior percepts. Imaging may also combine imagistic contents in new ways (as presumably, does a composer or a painter when creating a new piece). An artist, who can exploit these differences and their implications as part of what his/her piece suggests to an observer or a listener, can also, in such a case, consider several frames of reference in parallel. It may be supposed that the informed observer/listener will be able in turn to trace these various hints to contrastive approaches when exposed to the work.

Robert Gordon was right to stress that simulating was a basic way of performing practical reasoning.[11] But he apparently failed to observe that the crucial step for simulating mentally consists in the joint ability to represent *arbitrary* states of affairs *and* to reason *across frames of reference*. Simulating familiar, absent contexts is the main focus of pretend games. Children, around 2 years of age, exercise their practical reasoning ability in putting together counterfactuals and deriving relevant consequences in a world close to the real one (Leslie 1987). It should be noted, however, that the frame of reference of story telling or of pretend games is restricted to the isolated situation considered. There is only one world being referred to when a boy pretends being an Indian warrior, and the reasoning involves considering exclusively the properties that hold inside that world. Thus a playing child's way of simulating uses the same kind of capacity as when remembering familiar contexts (both kinds of memory involve representing only one context in a world). The only, but important difference is that the child engaged in pretend-play calls this representation actively, rather than in a passive associative mode. In pretend-play, some properties of the props are actively disengaged while others are retained (a banana does not hold as a fruit, but as an object shaped as a telephone). Fiction also

has the property of helping the child to project himself in a situation including a number of arbitrary properties which will be used as a basis for inferences – those being constitutive of the understanding of the story. There is no transparency of psychological attitudes at this stage in phylogeny or in human development: mammals, primates and young children exercise pretence without being able to fully understand what pretending involves. They know how to pretend without necessarily knowing that pretending conceptually imply believing.[12]

Exercising this imaginative ability in story understanding or in pretending however does not amount yet to a full understanding of pretending as a mental state. For such an understanding would require, in the view I defend, the further ability *to move within and without the simulated situation*.

In a primary stage of simulation, as found in pretend-play, no explicit representation of the world as it truly is, nor of the contrastive claims of truth made in the real world and in the pretended world, need be involved. This interpretation of pretend-play differs from a view in which make-belief presupposes that, in every episode of pretending, there is a set of beliefs incompatible with the pretended situation, which the pretender must explicitly rule out while she pretends. Evans (1982) for example analyses make-believe as involving both engaging in pretence *and* suppressing disbelief. In that analysis, simulating that *F has G* (when as a matter of fact, F does not have G) implies that one believes that "F does not have G" and *that one silences that belief*. As a consequence of this view, thinking counterfactually in pretence involves representing two worlds: the world of the game, and the world as it independently is. It also requires having a set of rules (a rule of incorporation and a recursive rule) establishing which truths from the real world still hold within the pretence.

In the present analysis, simulating that *F has G* does not *initially* presuppose that the subject believes that F does not have G. Weak counterfactual thought, therefore, consists in simulating a situation in which F does not have G, without representing explicitly whether doing so violates the rule for correctly representing F. In such a case, a canonical context for acting is presented as satisfied, using partial cues and ignoring those cues that do not match. The key element in pretending seems to consist in highlighting the relevance of actions, gestures and social meaning as determining a shared canonical context, to the detriment of various perceptual contents. The canonical gestures and their targets determine saliences in the perceptual scene. Pretending that one is in a car might be obtained by imaging a car and producing some auditory signal for the car's engine. A child pretending to drive a car does not need to

actively combat the belief that he is not actually sitting in a car. The pretence develops from the simple fact of imaging the dynamical properties associated to driving.

This analysis is consonant with developmental evidence. In pretend play, as well as in understanding narratives, children have initial difficulty at maintaining the scripts separate: they tend to take the products of their imagination as real, in particular when they are afraid by them.[13] They may also infer that some properties in the world have been changed as a result of the events in the fiction. Children finally learn that the *world of the game* and the *world in which the game occurs* are different, and license separate kinds of deductions. It is of course not clear how they learn this (to know how to respond to this would amount to having accounted for theory of mind acquisition). It may be through continuous practice, with the help of a more and more effective working memory. Alternatively, it may be through the help of a specific innate module coming to fruition around 3.

What is suggested is that there is more in strong counterfactual make-believing than "simple" pretence, and that the difference can be understood in terms of the frames of reference involved. This ability has been described, in Recanati's apparatus, by the distinction between *exercising* a simulation and *exploiting it.* Although we need not accept the theory of projection/retrojection defended by Recanati, the contrast drawn by Recanati is indeed close to the present view on strong counterfactual reasoning.[14]

If we now examine the structure of the specific type of simulation involved in psychological attribution, it appears that the concept of a frame of reference is essential in understanding exactly what it is to embed an ascent routine in another. What Gordon suggested is that, by successively engaging in a number of projections, one becomes able to grasp the notion of a mental location. The problem we identified above is that nothing constrains the simulator to entertain the notion of a causal role for belief. She might as well reason, as chimps probably do in their own modes, that agents have varying dispositions correlated with the behavioural and other physical cues available in the environment.

This is no longer possible in the case of the kind of embedded counterfactual reasoning that we are presently describing. In the exercising mode, the simulator must first project into a situation in order to draw various "*internal*" inferences from the properties that it contains. In the exploiting mode, the simulator must examine the external consequences of the projected situation for the real world (or for some other counterfactual world). For example, in a task of unexpected transfer, the child must first represent the counterfactual situa-

tion CS according to which the object did not change its location (from O to O′). She must then draw the consequences of this situation for an action relevant in CS (fetch the object in O), and finally infer from this the consequences for the real world: the simulated agent will go to O, even though the object is now in O′.

What may be an obstacle to understanding the role of exploitation in false belief tasks is that spatial properties are easily traceable from one world to the next: in the allocentric mode, locations can be determined absolutely. The unexpected transfer task relies on the ability to determine absolute locations, and memorise which locations have been salient in a past, now counterfactual situation. In such a task, the extra step from examining a counterfactual world to drawing consequences in the real world may well appear to be a useless sophistication.

Cognitive lying offers a finer-grained way of exploring how a subject exploits a simulation. The interest of this kind of ability is that it cannot rely on spatial memory alone. Let us suppose that Jean murdered his wife in their Parisian flat, but claims that he was away from Paris at the time of the murder. In order to make his lie accepted as a truth, the simulator must engage himself (through a narration or some kind of staging) into a counterfactual situation such that the desired properties follow: he was away watching an exhibit in London. But he also must compare the world of his pretence with the real world, in order to revise, among the many properties derivable in both worlds that are incompatible, those that can be found not to hold in the real world. Inter alia, all the signs of his presence in Paris must be erased, the Channel ferry or other public transportation systems must have been working, the exhibit must be on in London that day, etc. To be successful, the liar must both exercise a simulation of the fictional world he projects into, and exploit the real consequences of his fictional world, through maintaining a maximal coherence across both worlds. He will have to use the kinds of rules which Evans spelled out for make-believing: new facts can be incorporated in the lie as long as they do not contradict any statement (falsely) presented as true.

Exploiting a simulation thus first requires establishing in thought two types of representations: one of some pretended situation and the other of a model of the real world as it appears to the simulator. Then the simulator has to reason across these models. He has to evaluate the non-actual situation on the background of the real world, to derive which facts are or can be made compatible and which facts are incompatible (ignoring the property that forms the core of the lie). This step may be seen as a truly embedded simulation, or as two embedded uses of the same simulation. The question can now be raised

again of how this more complex kind of simulation allows an agent to acquire psychological concepts.

Complex simulation and psychological concept acquisition

In the present approach, a practical understanding of the concept of belief is supposed to be constituted by the capacity of evaluating the content of one simulation through the content of another. A subject who exploits simulation in the sense indicated has to draw inferences at least at two levels, allowing her to determine what is wrong in the simulation 1 compared to a simulation 2. Is the concept of belief, with its inferential structure and its associated truth-value, practically grasped as an outcome of the kind of embedded counterfactual reasoning described above? The practical know-how involved consists in exploiting the simulated perspective for one's own sake; i.e. deriving consequences inside a simulation that, properly redescribed, are relevant in the real world. This view is in agreement with a prediction of evolutionary psychology that psychological concept acquisition is part of the arms race driving selection of new capacities in phylogeny. Using what one understands as being conspecifics'views on things is of primary significance in order (i) to take advantage of what a foe knows or fails to know by manipulating him into improper actions; or (ii) to help a friend or a relative to acquire knowledge or revise his beliefs in order to act rationally.

There are two obvious questions to raise at this point. One is: why should the choice of the relevant assumptions in simulating some foreign perspective be any less circular in the present suggestion than in Gordon's one? The second is: how is a simulator able to grasp from his embedded simulation a notion not only of belief, but also of all the various psychological attitudes that can be entertained?

1. Simulation theory and vicious circularity

I will suggest that what appears as a vicious circularity might result from a perspective mistake. What could be termed "simple" simulation occurs only early in development (or in other primate species). Simple simulation is simulation conceived as a process reflecting only knowledge of physical facts and, for humans, linguistic expressions for occurrent belief states that are not fully understood as mental states. In other words, simple simulation is the kind of simulation of others that is exercised in the absence of psychological concepts.

Now it is clear that, at some point in development, psychological concepts and theories do help a simulator to adjust her simulations to the specific cases in which the projected situation departs on certain assumptions from the corresponding closest real world situation. On the other hand, it was claimed above that the really important step in simulation occurs when the situation to be simulated is not of a familiar kind, i.e. does not coincide with the memory of a past event experienced by the simulator. The difficult question is therefore the following: Can *this* kind of simulation – using both a counterfactual situation and another world of reference to which the former is compared – get off the ground without using psychological concepts?

Fortunately, there is a class of cases showing that it indeed is the case. The reason is obvious: the kind of simulation based on counterfactual reasoning that is necessary for mentalising is also needed in cases having nothing to do with psychological attribution. While planning a complex action, for example, one must be able to introduce new objects with a set of consequences (exercising simulation), and predict what the consequences on the planned course of action will be in a specific real-world context. For example, when planning to clean a well, one must consider the possibility that some toxic gas be present at the bottom and kill the cleaning human agent. This type of simulation involves various inferences *external* to the situation considered, through a process of exploitation of the simulation. It also needs a revision of the default assumptions that hold when cleaning an ordinary object. Such a revision of the situation does not need to appeal to any psychological knowledge.

If pure counterfactual simulation is available at the level of the physical world and exercised routinely in planning action, it may safely be concluded that this type of reasoning does not involve as a precondition a mastery of psychological concepts. In the case in which simulation is used to understand others and predict their behaviour through psychological attribution, one could suggest that the type of counterfactual reasoning involved is of the same kind, except that situations are indexed to definite individuals "whose situation it is". One could object here that "a fact to x" already involves an understanding of a mental access to x. But remember that a radical simulationist will export, so to speak, his own immediate epistemic relations to facts, appropriately restricted to specific contexts. The objection can therefore be countered by the routine ascent: I deal with P1, Peter deals with P2, and we form representations corresponding to P1 and P2. The notion of "dealing" as used by the simulator in fact draws on his own belief/desire structure, not necessarily on his belief/desire concepts. For the rival representations in question dot not need to be *represented as representations*, contrary to what theory-theorists would claim.[15]

They only need to be accessed demonstratively, as "this situation" versus "that situation".

Saying this does not commit the simulation theorist to the bold view that all there is to be known about belief can be gathered through this practical ability of embedded reasoning. The idea is that the practical reasoning ability is the foundation for understanding others, not that acquiring this ability through reasoning on various counterfactual contexts is *sufficient* to reliably predict others *in all cases*. Knowing the kinds of facts that are typically encountered by others is often not enough to simulate them; you also need to know what they specifically believe, desire, or fear to represent the situation as they see it and to predict accurately what they will do. It is thus plausible that the set of assumptions to be used in deriving the facts that hold as a consequence of some situations cannot in many cases be correctly established when the simulator does not fully master the concept of belief. Simulation theorists do not need however to deny this obvious fact. What they might suggest is that a child starts applying strong counterfactual simulation in a less-than-reliable way. An *imperfect* capacity for simulating others may be enough to get the practical know-how required for belief understanding started. In a situation such as the false belief task, for example, a simulating child only has to represent dynamically the goal and dispositions for action attached to the context as known, and those attached to the context before the transfer occurred.

But the preceding parallel of simulating others with counterfactual planning may give rise to a more serious objection. The latter focusses on the claim that reasoning across frames of reference, – exploiting the consequences of one frame in another – should *provide* the simulator with the practical knowledge of what it means to have a *false belief*. The objection[16] goes as follows. Let us imagine that A is able to reason across frames, and that he compares how things look to B with how things are. In such a scheme A can attribute a false belief to B, but this is a far cry from attributing false beliefs *to himself*. Considered as concepts, truth and falsity apply however *in a general way*. Such a generality might well fail to hold in the case of the kind simulation presently explored; even though simulating counterfactual situations involves a capacity to evaluate a possible world as not realised, there is no guarantee that a simulator will thereby be able to evaluate his/her own simulations and predictions as subject to error. Recognizing that one's *own* beliefs can be false, the objector concludes, involves more than reasoning across frames of reference; it involves understanding that beliefs are propositional attitudes, i.e that their contents can *constitutively* be true or false. Therefore our view on simulation plus coun-

terfactual reasoning is not better off than Gordon's radical simulationism. Both views are indeed circular.

Before attempting to answer this powerful objection, let us first remind what is at stake. The present paper, again, does not aim at defending the view that simulation might provide in and by itself all the necessary means for understanding other minds in all circumstances. What it does claim, however, is that a practical way of grasping and possibly putting to use another subject's representation of the world consists in *exploiting* counterfactual reasoning; i.e. drawing the consequences of a situation as simulated in the simulator's own world. As we will see below, this ability could be the underpinning of mental concept acquisition. Clearly, if it is shown that counterfactual reasoning of this kind can only be performed by a subject who is already mastering the concept of truth, then our claim will fall prey to circularity. Several authors indeed seem to accept the view that a subject only grasps the concept of truth when she possesses a theory of mind (Carruthers 1996; Papineau 2000).

Now one way of defeating the objection is to show that there is a conceptual difference between evaluating practically one's own embedded simulation and explicitly applying to it a meta-representational concept of truth. Accepting this conceptual difference does not lead us to expect however that it should be observable in empirically distinct predictive behaviors or rational strategies. Even though only a full-blown theory of mind can provide a subject with the fully general and explicit concept of truth, a simulator endowed with the kind of counterfactual reasoning described would indeed reach a coextensive practical knowledge. For truth is something that a simulator desires even though she does not master the concept of truth. A simulator wants her simulation to deliver predictions of which she can actually take advantage. As Papineau puts it, "If you act on true beliefs, you will generally get the results you want, but not if you act on false beliefs" (Papineau 2000: 201). In the case of simulation, this implicit, or practical desire for truth, needs to apply to one's own simulation. If you try and figure out what another subject is up to, you must be prepared to *revise* your simulation. A search for alternative explanations involves an implicit recognition that a given episode of simulation got things wrong. The distinction that the individual simulator needs to master at this point is of a successful/unsuccessful prediction.[17]

This argument from the very structure of revision in simulation can be completed by phylogenetic considerations on the origin of mentalising capacities. It is generally accepted that the selective pressure on more and more sophisticated mentalisers is based on the need to detect cheaters, liars, misinformed and malevolent communicators. Thus a reflexive dimension should

be immediately salient in counterfactual reasoning applied to others: knowing that one can be simulated and taken advantage of, by being offered a distorted view of reality is a crucial step in such reasoning. In order to understand that someone tries to lie to you, you must simulate yourself listening to a false indication, and acting on its basis in a way serving the other's ends. Thus a successful simulator is one that anticipates that he could be manipulated into misrepresenting things ; but again, this anticipation does not have to be explicitly understood in terms of truth or falsity, nor even in terms of representation; it is enough if the subject is able to imagine (or recall) the situation A where P holds, and contrast with situation B where P does not hold, with their contrasted affordances for the self or for others.

Let us summarize this point. Exercising simulation does not require a concept of simulation or representation, any more than believing requires having a concept of belief. Having the sense of a successful or unsuccessful simulation is the key to revising, but no concept of truth or of a true belief needs to be mastered. Finally the simulation of being possibly deceived is the essential reason why counterfactual simulation is being developed; this gives us a presumption, if not an argument yet, that a simulator is able to look at her own simulations as fallible.

The general solution of the problem of psychological concept application is that reasoning mentally involves a capacity to use perspectival information in strongly counterfactual cases. The latter ability is linked to executive properties of working memory, (maintaining active two rival representations for the same context) and not to psychological concept possession per se. Simulationists do not need to deny that this kind of reasoning should be considerably enhanced through linguistic expression. Symbol use will obviously help a child to build the relevant metarepresentations, to reason from them and to communicate rationally about them. This observation leads us to our second question.

2. Is a simulator able to acquire the various psychological concepts through simulation?

We saw earlier that Robert Gordon admits that simulation does not deliver a genuine, comprehending ascription of mental states, while nevertheless claiming that simulation may end up with "greater conceptual understanding" (1996: 17). We tried above to clarify in which way secondary simulation involves a specific kind of counterfactual reasoning. The question that needs to be raised at this point is to understand whether this *know-how* develops into a *know-that;* is a simulator able to grasp the general structure of reasoning

involved in mentalising, and to identify conceptually the various roles which propositional attitudes play in a subject's reasoning? How does a simulator manage to distinguish the various psychological concepts (believing, desiring, intending) that may be predicated of the various simulated contents?

As Heal remarks, Robert Gordon's view that simulating just implies putting oneself in a state similar to the target fails to provide the kind of psychological knowledge that a theory of mind of any variety should provide:

> If getting into states similar to those of others is to be of any use in psychological understanding, it must not merely occur; it must also be recognised for what it is. (Heal 1995: 44)

How then might we construct a simulator's understanding, say, that John *believes* that P, without positing circularly that the content of the simulator's pretence was [believing that P]?

A plausible suggestion for bridging the gap between the simulatory episode and psychological concept use might consist in invoking a procedure accepted in other fields of concept acquisition, i.e. demonstrative ostension. Heal presents this strategy as a general solution for identifying both the content of a simulation episode and the kind of simulation instantiated in this episode. What subject A simulates is that subject B entertains *this content.* A also actually simulates that P *in this way* (*believing*) rather than *in that way* (*desiring* or *fearing* that P). This solution however presents some problems of its own, (on which we will not expand here) having to do with the fact that simulating as an experience (referred to demonstratively) is not itself a belief experience (but rather a pretend-belief experience). Therefore pointing to a simulating episode will not allow pinning down a belief episode, but at best a pretend-belief one. At best, however: for when engaging under the pretend mode in a given situation, a simulator will normally have a variety of attitudes, from pretend-belief to pretend-desire and pretend-fear, not to mention various pretend-emotions. An additional difficulty is that the notion of a belief (or pretend-belief) *experience* is not clearly independent from the corresponding conceptual thought that one has a belief, or a pretend-belief. Thus the disintrication of contents and propositional attitudes linked to a given situation seems to be made no easier through the demonstrative procedure alone.

A more promising way of disintricating attitudes may consist in relying both on simulation as a reasoning procedure and on social learning. We saw earlier that the ascent routine, as usually understood, is only a way to parrot, rather than to express a judgement self-ascribing a belief. What makes it so is that the use of *believing* that is relevant in the routine ascent expresses only a

product of a primary simulation (or simple simulation) of a specific context. A "secondary" simulator however (able to perform strong simulation) transforms her prior understanding of the word "I believe", when developing her reasoning abilities. The question becomes one of articulating the know-how involved in reasoning *counterfactually* and the conceptual knowledge about belief and other psychological concepts.

A distinction made by Andrew Woodfield (Woodfield 1996, 1997) will help understand how such an articulation may be effected in the course of development. We will first present his view, presented as the author did in a way consonant with Theory Theory, and will adjust it later in terms consonant with simulation theory. The psychological concepts that are constitutive of a theory of mind, such as believing, desiring, and intending, should be seen as accessible public concepts, i.e. essentially communal norms. According to Woodfield, concepts in general are of this kind. They regulate human reasoning and communication in a way that does not presuppose their being explicitly grasped under a theoretical formulation:

> The player subscribes to the norms in advance of fully knowing what they are and without knowing in detail how to apply them, just as a club-member subscribes to the rules of the club without having properly read the rule-book. (Woodfield 1996: 91)

On the other hand, in his view, an individual thinker has a particular set of inter-related beliefs and know-how, however vague and inchoate, about any specific domain (folk-psychology as well as folk-physics, folk-biology, etc.). Her beliefs about belief, desire, and related concepts are a result of her particular experience, and thus differ from any other thinker's in some ways. These beliefs are theoretical in the sense that they go beyond observation. These individual theories, which Woodfield calls "conceptions", are "evanescent" personal constructions, in the sense that they are permanently revised and thus highly unstable, particularly so in the course of development between 2 and 6. A child may thus both have a common concept of belief in the sense that she understands sentences of a public language containing words expressing it, while possessing a different, idiosyncratic personal conception of belief (in which, for example, whatever true beliefs the self has are equally accessible to other selves). The important two points are first that, even in such a case, the child can be credited with *having the common concept of belief,* rather than with using some private or distorted concept; and second, that her conception of belief may become *more appropriate*, i.e. more *reliable*, when her individual conception gradually adjusts to the common norm.

Woodfield's approach can be accommodated within a simulationist framework. The core idea is that psychological concepts form a framework of socially accepted concepts and norms, to which every language user refers when communicating about psychological facts and events. In the course of development, a child needs two different types of information to apply these concepts in a reliable way to new sets of facts and events.

She must first have access to the relevant kind of linguistic exchange: "the ascent routine" is one of many standard ways of mastering a mentalistic language. Others of the same ilk are offered by parental guidance[18] in helping their children recognising emotions and various mental attitudes. The social character of these exposures to conceptual terms provides the content of concepts, although these concepts may only at first be used in very restricted kinds of inferences. In the present view, linguistic exchange provides a "proto-theory" for folk-psychology rather than a full-blown inferential structure. In the case of the mental, a crucial factor for building a psychological theory consists in what we called strongly counterfactual simulation.

The second kind of knowledge a mentalising child needs is thus the set of reasoning, representational and executive capacities that allow a thinker to use the social concepts in a fully general way. This kind of knowledge is simulatory and procedural. A child must know how to represent a type of situation across worlds and use this representation to make accurate predictions on how these worlds interconnect. Thanks to the second kind of learning, a child becomes able to become a creative mentaliser: she now has her own inferential capacities allowing her concept of belief to operate in new situations.[19]

As suggested above, both kinds of knowledge might actually be interdependent, in the sense of contributing to each other's development. Sentences retrieved from memory help simulate types of situations, and contribute to develop efficient execution. On the other hand, reasoning helps a child distinguish in which relevant aspects two linguistic descriptions of a situation type differ. Saying this does not amount to acknowledging that simulation does not have a structuring role in allowing a child to mentalise.

Conclusion

Let us summarise the solution offered above to the problem of psychological concept acquisition. This solution seems in its core to be in the spirit of Gordon's radical simulationism. The basic ingredients include again simulatory practice and ascent routine. Our expanded theory, however, contrasts two

kinds of simulations, and emphasizes the structuring role of counterfactual reasoning in embedding simulations and deriving facts stable across them. Furthermore, it completes ascent routines with social learning of folk-theoretical words.

When exposed to a public language, a child is provided with sentences relating desire, belief, and intention expressions. The various terms involved in a theory of mind are first grasped in a shallow and idiosyncratic way as a result of the child's developping mastery of public language. As we saw, such a linguistic mastery of social concept-terms does not amount yet to possessing a theory.

Independently from her linguistic ability in using ascent routines and echoing adult's talk about belief and desire, a child relies on her developing reasoning and executive memory capacities to anticipate others' doings, emotions and intentions. Strong counterfactual simulation seems to be the key ability for such an ability to anticipate. She may on this basis develop a personal conception of lying, being misinformed, and having weird desires. This personal conception will gradually be merged into the common, socially transmitted, folk-psychological theory.

The threat of vicious circularity seems thus to dissolve. For a subject does not need to know herself to know others. She rather needs to evaluate situations in a way both psychologically and rationally similar to other subjects. In the proposed view, as in Gordon's approach, psychological knowledge has nothing to do with an introspective self. The present view stresses further its social nature. Mastering a theory of mind cannot be reached independently of social interaction. But an individual can only benefit from social theory-learning if she is able in practice to represent counterfactual situations dynamically, and to extract from them stable properties.

Notes

1. See Heal (1995: 37), (1996: 75).

2. This may be a problem for a full grasp of the concept of belief, but it may not be an objection to a theory of folk psychological concepts. It seems plausible that people tend to deny spontaneously that a proposition that nobody believes or even considers as possibly true indeed refers to a fact.

3. See Hood (1995).

4. For a detailed discussion, see Proust (2000b).

5. Trevarthen (1999).

6. See Daprati et al. (1997), Jeannerod (1999), Proust (2000a).

7. Pragmatic information concerns the kind of action performed, whereas praxic information concerns the kind of movement performed when acting.

8. The role of counterfactual reasoning in theory of mind acquisition is also stressed in Riggs et al. (1998) and in Donald Peterson's contribution in this volume.

9. Leslie and German (1995: 129); Perner (1996).

10. For a more detailed presentation, see Proust (2002).

11. See also Heal (1996: 78).

12. For a similar view, see Heal (1995: 40).

13. Harris et al. (1991).

14. According to Recanati, *exercising* a simulation consists in accepting a set of premises as holding in a particular imagined situation, and drawing from them some relevant consequences inside the same imagined situation. Accepting the premises does not mean that the simulator must take them as *true.* He only *pretends* them to be true, or *imagines*, or *supposes* them to be true. Deriving further properties from the supposed ones does no longer belong to pure imagination. Such a conditional reasoning may be taken as involving typical relations within the real world or a subset of them. *Exploiting* a simulation allows considering the same situation both from the alien point of view in which we project ourselves (through exercising simulation) and from our own point of view. Exploiting simulation allows the simulator to consider the situation simulated not only as a set of meaningful relations between properties, but also as having certain relations with the actual world. For example, understanding Mallarmé's poems, requires both exercising and exploiting simulation, i.e. requires a parallel consideration of the situation depicted, and what belongs to the writing conventions. This is not a situation peculiar to literature. Understanding mentally another person requires both seeing things as she does and retrodicting the causes or predicting the real effects of the simulated situation in connection to the world as we see it. While exercising a simulation involves a projection within a situation, exploiting a simulation involves an embedding of the former situation within a real world perspective, a process which Recanati calls *reflection* (1999: 101).

15. In particular, Perner (1991).

16. I thank Jérôme Dokic for articulating this objection.

17. It will be noted that the appearance/reality distinction, as studied by Flavell et al. and many other developmental psychologists, does not coincide with the distinction between two stages in acquiring the concept of truth. Our notion of counterfactual simulation allows a simulator to grasp practically this contrast as one of differing simulated points of view.

18. Cf. the role of mimicry in typing mental states is hypothesised in Gergely and Watson (1998). See also Proust (2002).

19. Although we cannot develop this point here, reliability of simulation procedures for understanding others' mental states may be articulated in terms of the specific conditions for simulating: that the processes of thinking across subjects (simulator and simulatee) are the same or sufficiently similar ; that the context simulated is the same or sufficiently similar; that the knowledge of the context is the same or sufficiently similar; and finally, that the

simulator has the reasoning and executive memory capacities allowing him to reason across worlds in the way described above (see Heal 1995: 40).

References

Baron-Cohen, S. & Goodheart, F. (1994). The "seeing-leads-to-knowing" deficit in austim. *British Journal of Developmental Psychology, 12*, 397–401.

Biro, S. & Russell, J. (2001). The execution of arbitrary procedures by children with autism. *Development and Psychopathology*, *13*, 97–110.

Carruthers, P. & Smith, P. K. (Eds.). (1996). *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the Agent, an Investigation into Self-consciousness and Consciousness of the Action in Schizophrenic Patients. *Cognition*, *65*, 71–86.

Davies, M. & Stone, T. (Eds.). (1995). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.

Davies, M. & Stone, T. (Eds.). (1995). *Mental Simulation*. Oxford: Blackwell.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). Brain activity during observation of action. Influence of action content and subject's strategy. *Brain, 120*, 1763–1777.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Clarendon Press.

Gallese, V. & Goldman, A. (1998). Mirror neurons and the simulation theory of mind reading. *Trends in Neuroscience, 2* (12), 493–501.

Gergely, G. & Watson, J. S. (1998). Early Social-Emotional Development: Contingency Perception and the Social-Biofeedback Model. In P. Rochat (Ed.), *Early Social Cognition* (pp. 101–136). Hillsdale, NJ: Erlbaum.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language, 4,* 161–185.

Goldman, A. I. (1992). In defence of Simulation Theory. *Mind and Language, 7* (1–2), 104–119.

Goldman, A. I. (1993). The Psychology of Folk Psychology. *Behavioural and Brain Sciences, 16*, 15–28.

Goldman, A. I. (1995). Empathy, Mind and Morals. In M. Davies & T. Stone (Eds.), *Mental Simulation* (pp. 185–208). Oxford: Blackwell.

Goldman, A. I. (2001). Desire, Intention, and the Simulation Theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and Intentionality, Foundations of Social Cognition* (pp. 207–224). Cambridge: MIT Press.

Gordon, R. M. (1993). Self-ascription of belief and desire. *Behavioural and Brain Sciences, 16* (1), 45–46.

Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In M. Davies & T. Stone (Eds.), *Mental Simulation* (pp. 53–67). Oxford: Blackwell.

Gordon, R. M. (1996). 'Radical' simulationism. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 11–21). Cambridge: Cambridge University Press.

Harris, P. (1992). From simulation to Folk Psychology: The Case for Development. *Mind and Language, 7*, 120–144.

Harris, P. (1993). Pretending and Planning. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other Minds: Perspectives from Autism*. Oxford: Oxford University Press.

Harris, P. L., Brown, E., Marriott, C., Whittall, S., & Harmer, S. (1991). Monsters, ghosts and witches: testing the limits of the fantasy-reality distinction in young children. *British Journal of Developmental Psychology*, *9*, 105–123.

Heal, J. (1995). How to think about thinking. In M. Davies & T. Stone (Eds.), *Mental Simulation* (pp. 33–52). Oxford: Blackwell.

Heal, J. (1996). Simulation, theory, and content. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 75–89). Cambridge: Cambridge University Press.

Hood, B. (1995). Gravity rules for 2–4 year olds. *Cognitive Development, 10*, 375–386.

Jeannerod, M. (1999). The 25th Bartlett lecture; To act or not to act. Perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology, 52A* (1), 1–29.

Leslie, A. M. (1987). Pretence and representation, The origins of 'theory of mind'. *Psychological Review, 94,* 412–426.

Leslie, A. M. & Thaiss, L. (1992). Domain specificity in conceptual development. Neuropsychological evidence from autism. *Cognition*, *43*, 225–251.

Papineau, D. (2000). The evolution of knowledge. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the human mind, Modularity, language and meta-cognition* (pp. 170–206). Cambridge: Cambridge University Press.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge: MIT Press.

Perner, J. (1994). The necessity and impossibility of simulation. *Proceedings of the British Academy*, 29–44.

Perner, J. (1996). Arguments for a simulation-theory mix. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories mind* (pp. 90–104). Cambridge: Cambridge University Press.

Povinelli, D. (1996). Chimpanzee theory of mind? The long road to strong inference. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 293–329). Cambridge: Cambridge University Press.

Proust, J. (2000a). Awareness of Agency: Three Levels of Analysis. In T. Metzinger (Ed.), *The Neural Correlates of Consciousness* (pp. 307–324). Cambridge: MIT Press.

Proust, J. (2000b). Can non-human primates read minds? *Philosophical Topics, 27* (1) 203–232.

Proust, J. (2002). Imitation et agentivité. In J. Nadel & J. Decety (Eds.), *Imitation, Représentations motrices et intentionnalité*. Paris: Presses Universitaires de France.

Quine, W. V. O. (1960). *Word and Object*. Cambridge: MIT Press.

Recanati, F. (1999). *Oratio Obliqua, oratio recta*. Cambridge: MIT Press.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, *13*, 73–90.

Strawson, P. F. (1959). *Individuals*. London: Methuen and Co.

Trevarthen, C., Kikkinaki, T., & Fiamenghi, G. A. (1999). What infants' imitations communicate: With mothers, with fathers and with peers. In J. Nadel & G. Butterworth (Eds.), *Imitation in infancy* (pp. 127–185). Cambridge: Cambridge University Press.

Tulving, E. (1983). *Elements of Episodic Memory*. New York: Oxford University Press.
Vygostsky, L. S. (1978). *Mind in society*. Cambridge: Harvard University Press.
Woodfield, A. (1996). Which theoretical concepts do children use? *Philosophical Papers, XXV* (1), 1–20.
Woodfield, A. (1997). Social externalism and conceptual diversity. In J. Preston (Ed.), *Thought and Language* (pp. 77–102). Cambridge: Cambridge University Press.
Zaitchik, D. (1990). When representations conflict with reality: The pre-schooler's problem with false belief and 'false' photographs. *Cognition, 35,* 41–68.

Reply to Joëlle Proust

Pascal Ludwig
Université de Rennes

I shall begin this paper by rehearsing Wittgenstein's important distinction between two uses of the pronoun "I": as object and as subject. Used as object, it behaves like a referential expression. Its referent in context can be identified by observation, or with the help of a description. A different use appears in sentences like "I feel pain", "I am relieved", "I see a napkin on the table" or "I think the lecture is starting now". According to Wittgenstein, we do not refer to ourselves in such sentences. This claim is controversial,[1] but we do not need to endorse it to notice that there is something right in Wittgenstein's distinction. In its uses as subject, the pronoun "I" is immune to error as far as the identification of the referent is concerned. As Wittgenstein writes, "it is impossible that in making the statement "I have toothache" I should have mistaken another person for myself, as it is to moan with pain by mistake, having mistaken someone else for me." I do not need to exercise a concept of myself, be it indexical, recognitional or descriptive, in order to self-ascribe properties like "being in pain". As I understand it, Wittgenstein's important insight is that at least some first-person thoughts are *from* a perspective, or from a point of view, rather than *of* a perspective or a point of view. Drawing on this insight, many philosophers have insisted that there are two kinds of representations: egocentric representations and objective representations.[2] Egocentric representations are sensitive to the context in which they are exercised, whereas objective representations are not. We need egocentric representations, these philosophers insist, because we are agentive cognitive systems.[3]

According to Simulation Theory, the psychological capacity of predicting others' behaviour is grounded on imaginative skills. To predict successfully what others will do, a child must be able to imaginatively identify with someone else and imagine a counterfactual situation. There are, however, weak and strong forms of this general assumption. According to weak forms of simula-

tionism, simulating is analogue to modelising in science. Mental concepts, in such a view, are grasped through the grasp of a theory, that is, a set of general predictive principles involving those concepts. Nobody could predict his peers' courses of action without knowing, at least tacitly, such principles. This does not imply that mental theoreticians should not use imagination when they apply those principles, or when they draw inferences form them: it is often useful to use our capacity to simulate in order to apply general principles, or to reason. This is true not only of the mental domain, but also of many other theoretical areas. Consider theoretical physics, for instance. In order to make sense out of complex data, and to apply general principles to them, physicists often construct bridge models. In doing so, they imagine that the data follow certain patterns that can be described by simple equations. In other words, they imagine that the data exhibit certain regularities, even if they do not really do so. This is pretense, just as the ability to project oneself in someone else's situation is pretense.

Do agents, however, really use their own minds as models of the psychological processes they try to predict? If this analogy was correct, self-observation would play a central role in simulation. Following Robert Gordon, Joëlle Proust argues to the effect that this is not so. Simulation, she insists, occurs through an imaginative transformation into the other, which allows the simulator to see the situation that the simulatee faces as she faces it. The agent undergoing a simulation process should not be understood as introspecting her own psychology, as scrutinizing her own states of mind in a reflexive way, but rather as looking at the environment as it is from the perspective she imagines to be projected in. The key notion here may be called *projection* or *transference*: simulating someone amounts to projecting oneself in her situation.

This central idea may seem obscure at first. What does "imagining being someone else" mean? Is it indeed conceptually possible to imagine being someone else? Joëlle Proust addresses this issue in an illuminating way. She claims that transference into another situation makes sense as soon as it is interpreted as a context-driven process, exploiting egocentric, or perspectival, mental representations. I agree with her analysis. In the first part of my comment, I shall put forward further arguments supporting her interpretation.

Even if one grants that transference is conceptually possible, another aspect of radical simulationism is controversial. According to Gordon, transference should be seen, in an important sense, as a basis for mental concept acquisition. This is an important point of divergence between Gordon's and Goldman's interpretations of simulation. According to Goldman, mastering mental concepts is a prerequisite for simulating others.[4] Gordon, on the other hand,

insists that early and simple forms of simulation do not require such a mastery. Proust proposes a critical assessment of this debate. With some qualifications, she sides with Gordon. In the second part of my comment, I shall present what I think is the main difficulty for Gordon's position and I shall discuss Proust's view on the matter.

1. Two grades of simulative involvement

It is intuitively tempting to characterise simulation in terms of projection. To borrow Quine's words about indirect quotation, simulating someone's behaviour amounts to projecting oneself into her mind, in an essentially dramatic way. Sometimes, simulationists also talk of imagination instead of projection. Thus, Goldman writes that "introspectively, it seems as if we often try to predict others' behaviour – or predict their (mental) choices – by imagining ourselves in their shoes and determining what we would choose to do".[5]

Do we have to take seriously the idea of projecting oneself into "someone else's situation", an expression that would sound like a mere metaphor to many? Proust convicingly claims that we do. She insists that we have a peculiar access to others' behaviours. We do not perceive these behaviours as mere bodily movements, but as actions intentionally directed toward goals. In her paper, Proust mentions some fascinating empirical results that tend to show that actions perceived by a subject are classified by their goals, and encoded in the same format as planned actions are. She interprets these results as showing that an observer is always involved empathically in one of her peers' behaviour: as she puts it in substance, seeing an action seems to constitutively involve a disposition to act in a similar way. I find Proust's interpretation of the empirical data she examines illuminating. A schema like the following one seems indeed quite plausible:

Behaviour –> Activation of motor image –> Off-line activation of –> Predictions
(encoded in a goal-specific way) motor routines

We have to notice, however, that a kind of homonymy is lurking behind such a framework. We started with a notion of sympathetic involvement, conceived as an imaginative projection into someone else's situation; what we have here are relations between sub-personal systems. Now, imaginative involvement consists in personal, intentional activity. When I imagine to be in a fictional or counterfactual situation, I certainly engage into intentional, conscious, think-

ing, whereas I certainly do not choose in any way to activate certain areas of my motor cortex in order to mirror one of my conspecifics' course of action.

We should therefore agree that there are at least two grades of simulative engagement into someone else's practical situation. The first grade involves the activation of sub-personal systems and of motor scripts and routines. It does not presuppose any conscious imaginative involvement into someone else's situation, or the entertainment of assumptions concerning her mental life. But it presupposes that the practical situation simulated be observable for the simulator. Proust, nevertheless, acknowledges the existence and the importance of a second grade of simulative involvement, which she calls "mental simulation". This grade involves an imaginative activity. A subject "mentally simulates" one of her peers if she imagines being in her shoes, or, more literally, *being her*.

2. Transference

According to the radical interpretation of simulationism, the expression "imagining being someone else" has to be taken very seriously. As Proust emphasises, Gordon's use of Gareth Evans' concept of an "ascent routine" is central here. Evans' main motive for introducing this concept is the conviction that mental self-ascriptions are not based on introspection or self-observation. There is an important distinction, according to Evans, between deciding for myself whether or not a certain state of affairs holds *in the world*, and deciding whether or not *I* am in a given mental state. In order to answer the first question, I must observe the world. It would be a Cartesian illusion, however, to conclude that I have to look into *myself* to decide whether or not I believe, for instance, that it is raining now:

> If someone asks me "Do you think there is going to be a third world war?", I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. (Evans 1982: 224)

One of Gordon's most original contributions to the Theory-Simulation debate consists in his extension of Evans' "ascent routine". Suppose, Gordon claims, that an agent may be able to imagine *being someone else* – say, that Marie may be able to imagine being Pierre. Let us call "transference" this ability to project oneself into someone else's situation. It should be possible to apply the ascent

routine into the shifted new context the agent imagines being in. Let us consider, for instance, a real situation involving two children, Mary and Jane, playing with a Mickey Mouse doll.[6] How could Jane decide whether or not Mary believes that Mickey Mouse has a tail? According to Gordon, a correct explanation of Jane's mental process during simulation involves both transference and an application of the ascent routine. Here is an outline of this process:

1. Jane first projects herself into Mary's situation (transference).
2. She then tries to answer the question: "Does Mickey Mouse have a tail in the situation I am in?" (ascent routine).
3. As a result of transference, "the situation I am in" denotes the shifted context, that is, Mary's epistemic situation.
4. It follows that the application of the ascent routine yields the result it would have given if it had been run by Mary.

Gordon insists on the importance of the first step in this process, transference. Thanks to transference, the child is able to project herself in a shifted context, and to adopt someone else's epistemic point of view in imagination. He explicitly says that the denotation of the indexical pronoun "I" is shifted under this imaginative activity:

> (...) when Jane, in the role of Mary, generates the response, "Yes I (Mary) believe Mickey Mouse has a tail", she will understand the "I" to refer to that particular individual, Mary; and when, simulating another individual, Joe, she generates the response, "Yes, I (Joe) believe Mickey Mouse has a tail", she will understand the "I" to refer to Joe. (Gordon 1996: 18)

As I emphasised in the introduction of the present paper, not all simulation accounts share this assumption. In (Gordon 1992), Gordon contrasts his own interpretation of simulative processes with another interpretation, which he calls the "model model". The difference between these interpretations hinges on the possibility of reference shifting for indexical expressions like "I" or "now". According to the "model model", a simulator has to look into *himself* in order to draw useful conclusions about others; simulation essentially amounts to self-observation. To this interpretation, Gordon replies that simulation involves total transference: if an agent imagines being Napoleon, he imagines being in a situation in which the pronoun "I" denotes Napoleon instead of himself:

> What is relevant is to ask, within a pretend context in which I simulate being Napoleon or being Napoleonic, "What shall I do now?" The crucial difference is that in such a context "I" and "now" do not have their ordinary references: they do not refer, respectively, to the distinct individual doing the simulating

> and the distinct time he is doing it. (...) what remains constant is not the reference of personal pronouns (or other "essential indexicals"), but rather the special ties they have because our perceptions, memories, actions, and emotions are keyed to our egocentric map.
>
> (Gordon 1992, in Davies & Stone 1995: 119)

The possibility of such reference-shifting in make-believe contexts has been much discussed in the literature on imagination.[7] Richard Wollheim and Zeno Vendler, like Gordon, defend the possibility of imagining being someone else.[8] Others, like Kendall Walton and David Nozick, are more skeptical about such a possibility.[9]

I side with Proust and Gordon in this debate. My first reason for doing so is negative: I think that the reasons usually advanced against full transference are bad. The main argument against it relies on a conception of the reference of indexical expressions inspired by the work of David Kaplan.[10] According to Kaplan's view, these expressions are directly referential. This means that their denotations are never shifted by any intensional operator. There is a contrast, to this extent, between descriptions and indexicals: the denotation of a description can always be shifted when the description falls under the scope of an intensional operator, whereas the denotation of an indexical remains unchanged. Compare (1) and (2):

(1) The president could have been a communist.

(2) I could have been a communist.

In one reading, (1) is true if the president that could have been elected – that is, the president that has been elected in a possible world – is a communist. There is no such reading for (2): the occurrence of "I" must refer to myself; it cannot denote an individual, different from PL, that "I" could have been.

What conclusion should we draw from this fact? It seems clear that the denotations of certain expressions cannot be shifted by intensional verbs. This certainly does not mean that one cannot pretend to be in a context in which the denotation of such an expression *is* shifted. Consider, for instance, a fictional name like "Sherlock Holmes". This name has no referent, since Sherlock Holmes does not exist. Since it is a directly referential expression, or at least a rigid designator, there is no possible world in which it has a denotation. This does not prevent us from using the name *as if* it had a denotation, when we make-believe that the famous detective *does* exist. But if I can use a proper name as if its denotation was shifted, I do not see any reason why I could not use the personal pronoun "I" in a similar way. This is what happens on stage,

when an actor pretends being someone else. Hence, direct reference theory is consistent with the possibility of transference.

There is also a positive reason to endorse Gordon and Proust's claim. According to their position, simulating someone implies imagining being her, in a subjective, perspectival, way: that is, putting oneself in her shoes, in order to see the world from her point of view. The intentional contents involved in this imaginative activity are supposed to be first-personal, indexical thoughts.

Does such an activity of subjective make-believe really exist? It seems it does. Zeno Vendler emphasises a distinction between a subjective and an objective way one may imagine oneself doing something. Let us suppose that I am asked first to imagine swimming, and then to imagine *myself* swimming. In order to achieve the first imaginative project, I must take the subjective point of view of a swimmer, including his qualitative experiences: "the cold, the salty taste, the tug of the current, and so forth", as Vendler says.[11] There is a very different imaginative project, however, which consists in imagining *myself* swimming in an objective way. I may for instance imagine being in a boat and watching myself floating in the sea through a porthole. Vendler's conclusion is worth quoting:

> I shall call the act of imagination involved in the first exercise subjective, and the one involved in the second objective. As the swimming example shows, the second kind consists in the inner representation of one's body (or voice) from a certain perspective. This task is not different from the representation of other bodies; after all, I can imagine you floating in the ocean with equal ease. The other kind does not call for the representation of my body from a distinct point of view; it merely consists in evoking the experiences I would have if I were in some situation or other. (Vendler 1984: 43)

Vendler does not stop here. He further argues that objective imagination may be reduced to subjective imagination. This is a convincing point for anyone taking phenomenology seriously. When I objectively imagine myself swimming, I do it from a point of view, or a perspective. The very notion of imagining a situation from no point of view at all seems absurd.[12] It thus seems that subjective imagination is logically prior to objective imagination. This confirms, at least *prima facie*, the radical simulationist position: after all, if imagining always amounts to imagining being in a certain kind of situation, why should it be especially difficult to imagine endorsing someone else's epistemic perspective?

Suppose I am standing on the top of a cliff, watching my friend Pierre swimming one hundred feet below. Can I imagine being Pierre? I certainly can imagine that our positions, and therefore our perspectives, are switched,

since I can imagine swimming in the sea, instead of watching someone doing it. As long as "imagining being X" just means "imagining being in a certain kind of situation, and having certain kinds of experiences", the notion is rather unproblematic. Things become much more complicated as soon as we take epistemic properties into account. Imagining being Pierre does not only mean imagining being in his physical position, imagining perceiving things from his point of view, feeling the way he feels, etc. Another mind is a complex of bodily states, experiences, feelings, but also of memories, beliefs, desires, and complex intentions. Here, to borrow Vendler's words, we do "reach the limit of the power of subjective imagination": imagining being Pierre means imagining having Pierre's knowledge and Pierre's desires and intentions – at least the relevant ones.[13]

3. Imagining having epistemic properties

At this point, radical simulationism must face a traditional, but nevertheless powerful, objection. I quote Proust's statement of this problem: "A first well-known problem with the view that a simulator only needs to use the information contained in some perspectival situation is that some mastery of epistemic notions must be present for a subject to select, during deduction, those premises which are accepted by the simulatee". My focus, until the end of the comment, will be on this problem.

Proust's main argument consists in a plea for a shift of perspective. She thinks that the discussions have focused too much on complex examples, taken from our experience as adults. She recommends us to start, instead, from thought activities such as simple counterfactual reasoning, solitary imagination, games of make-believe, and fiction understanding. What she tries to show – convincingly, I think – is that those activities require complex cognitive achievements. The "ability to represent and simulate arbitrary states of affairs", first, but also an ability both to draw inferences in the context of such a simulation, and to use the conclusions of those inferences in the real world.

I do not want to enter into any detail here, because I think that Proust's argument to the effect that a mastery of those capacities is necessary in order to be successful as a mental simulator is convincing. This should not be the end of the story, however. Even if we grant that a mastery of those capacities is *necessary* for mental simulation, we do not need to grant that it is *sufficient for it*.

Let us contrast activities such as imagination in games of make-believe with full-blown mental simulation. In games of make-believe, the information about the situation the person has to imagine to be in is given to the child. Mum reads a story to her child, who has just to accept as true the sentences read in order to follow the rules of the pretense play. In symbolic games, children sometimes have to draw simple inferences. For instance, a child observes two of her mates playing with a puppy. To enter the game and be involved into the imagined situation, she has to recognise the situation type she has to make-believe to be in from the observed situation. Such recognition clearly presupposes an inference. But the important point is that this inference itself does not presuppose any attitude ascription from the child. The child does not have to master mental concepts, such as belief and desire, in order to be able to recognise a puppy as a prop for a certain type of make-believe games.

Consider, on the other hand, a typical situation involving "mental simulation". Suppose that I must meet one of my friends in Paris-Charles-de-Gaulle airport, but that we did not fix any meeting point. Where are we going to meet? In order to answer this question, I must determine where my friend is likely to go inside the airport. Intuitively, this is a typical mental simulation situation. To solve it, the best recipe is certainly to put myself in his shoes, and to observe what is going on.

It seems that assumptions concerning his mental states are crucial here. Does he know that we are supposed to meet? If he does not, he is likely to go directly toward the taxi station, and I am likely to find him there. But if he does, I must answer further questions. For instance, does he know that there is a meeting point at the airport? Does he know how to find it? Does he believe that I am waiting for him at the airport, and not at my Parisian flat? Of course, the answers to these questions are crucial to the process by which I determine in what kind of situation he is. But one cannot even formulate the questions without mastering mental concepts.

The upshot of all this is that it is not enough to show that a command of the non-epistemic capacities singled out by Proust is sufficient for some activities involving imaginative engagement to conclude that they are sufficient for the very specific activity that she calls "mental simulation". The really difficult question she must face is therefore the following: how does mental simulation emerge from simpler activities such as counterfactual reasoning, solitary imagination, or make-belief?

Proust is fully aware that this second question is also a difficult one, and she develops an interesting answer to it. Here is this solution, as I understand it:

> Of course, mental simulation has some very special features. In mental simulation, we imagine being in a situation that is "indexed to a definite individual" whose situation it is. That is, we project ourselves in a specific individual's situation. In the context of such pretense, it is possible to use the routine ascent, and to pretend-believe what the other really believes. We do not need to represent the other's mental states in order to pretend-believe something, however. We just have to represent the content of his states, and to pretend-believe them. Therefore, even mental simulation does not presuppose any mastery of mental concepts.

Let us apply this idea to my airport example. Here are the steps of the mental process I should go through according to Proust's solution:

1. Projection in my friend's situation by imaginary identification
2. Pretend-believing that I must meet someone waiting for me at a certain place in Paris
3. Pretend-knowing that I must take a taxi to this address
4. Classifying reaction
5. De-identifying
6. Attributing reactions to my friend and predicting his behaviour

I agree that steps 2 to 6 do not presuppose any mastery of mental concepts. I do not need, for instance, to represent my friend's belief state in order to pretend-believe that I must meet someone waiting for me at a certain place in Paris. But I claim that the first step in the process does presuppose such mastery.

In order to see why, let us analyse this first step. It seems clear that to project oneself into someone else's situation just means to imagine being her. But how could one imagine being someone else without making assumptions about her mental states? According to Leibniz's law, two entities are identical if and only if they share all their properties. So what we must make-believe, in a mental simulation, is to instantiate all the relevant properties of the person whose behaviour we intend to predict. But we insisted above that this set of relevant properties contains intentional, or mental, ones. If my aim is to predict the behaviour of my friend, imagining that my hair is brown, or that my eyes are black, is irrelevant. On the other hand, imagining believing that we must meet at my Parisian flat, for instance, is crucial. So mental properties have a central place in mental simulation. It is difficult to see, however, how a subject could make any assumption about the mental properties of one of his conspecifics without possessing and mastering mental concepts.

If we make this presupposition explicit, we get the following mental process:

1. Assuming that my friend has a certain set of relevant attitudes
2. Identifying with him by pretending to have those attitudes
3. Pretend-believing that I must meet someone waiting for me at a certain place in Paris
4. Pretend-knowing that I must take a taxi to this address
5. Classifying reaction
6. De-identifying
7. Attributing reaction to my friend and predicting his behaviour

In a way, I think that Proust herself is not all that far from reaching a similar conclusion, since she grants it to be clear that "at some point in development, psychological concepts and theories do help the simulator to adjust his simulation to the specific cases in which the projected situation departs on certain assumptions from the corresponding closest real world assumptions." But still, she seems to claim that the capacity to entertain such assumptions emerges from more basic simulations. The main question I want to ask her to conclude is how exactly this emergence takes place, given the specificity of mental simulation.

Notes

1. See Brinck (1997: Chap. 2) for a presentation of this controversy.

2. See Corazza (1994) and Moore (1997) for a detailed discussion of this distinction. Egocentric representations are sometimes called "indexical" or "perspectival."

3. See Perry (1993) and Mellor (1989).

4. See Goldman (1989), reprinted in Davies and Stone (1995).

5. See Goldman (1989), reprinted in Davies and Stone (1995: 82).

6. See Gordon (1996: 18).

7. See Feagin (1996: 85–86) for a concise review of the relevant literature.

8. See Wollheim (1984: 75–76) and Vendler (1984). Here is Wollheim's conclusion: "that I can centrally imagine being someone other than myself is a fact of major importance, and it is a fact of philosophy, not just a point of terminology." To support this conclusion, Wollheim remarks that imagining myself being someone else – let us say, being Napoleon – leaves it open to me to imagine myself brought face to face with him. That is, I may imagine being Napoleon, and imagine being introduced to Pascal Ludwig.

9. See Walton (1990: 31–33) and Nozick (1981: 636–638).

10. See Kaplan (1989). Kaplan calls "monster" any operator that would operate on context, and that would therefore be able to switch the denotation of indexicals falling into its scope. The existence of such operators, according to him, is conceptually possible; but, as a matter

of fact, they do not exist in English: "I am not saying we could not construct a language with such operators, just that English is not one. And such operators could not be added to it" (Kaplan 1989: 510).

11. See Vendler (1984: 43).

12. Peacocke (1985) defends the same point: he claims that imagining always amounts to imagining being in a certain mental state, from the inside, or from a subjective perspective.

13. See Vendler (1984: 53).

References

Brinck, I. (1997). *The Indexical "I": The First Person in Thought and Language*. Dordrecht: Kluwer.

Carruthers, P. & Smith, P. K. (1996). *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Corazza, E. (1994). Perspectival Thoughts and Psychological Generalizations. *Dialectica, 48*.

Davies, M. & Stone, T. (Eds.). (1995). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.

Davies, M. & Stone, T. (Eds.). (1995). *Mental Simulation*. Oxford: Blackwell.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Feagin, S. L. (1996). *Reading with Feeling*. Ithaca and London: Cornell University Press.

Goldman, A. I. (1989). Interpretation psychologized. *Mind and Language, 4*, 161–185.

Gordon, R. M. (1992). The Simulation Theory: Objections and Misconceptions. *Mind and Language, 7*, 30–32. Reprinted in Davies & Stone (1995) (pp. 100–122).

Gordon, R. M. (1996). "Radical" simulationism. In Carruthers & Smith (1996), (pp. 11–21).

Kaplan, D. (1989). Demonstratives. In Almog, Wettstein & Perry (Eds.), *Themes from Kaplan* (pp. 481–563). New York: Oxford University Press.

Mellor, H. (1989). I and Now. *Proceedings of the Aristotelian Society, 89*, 79–94.

Moore, A. W. (1997). *Points of View*. Oxford: Clarendon Press.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge: Harvard University Press.

Peacocke, C. (1985). Imagination, Experience, and Possibility: A Berkeleyan View Defended. In J. Foster & Robinson H. (Eds.), *Essays on Berkeley*. Oxford: Clarendon Press.

Vendler, Z. (1984). *The Matter of Minds*. New York: Oxford University Press.

Walton, K. (1990). *Mimesis as Make-Believe*. Cambridge, MA and London: Harvard University Press.

Wittgenstein, L. (1958). *The Blue and Brown Books*. Oxford: Blackwell.

Wollheim, R. (1984). *The Thread of Life*. New Haven and London: Yale University Press.

Chapter 9

Joint attention and simulation

John Campbell
University of Oxford

1. Understanding someone else's perceptual demonstratives

My topic in this talk is the way in which we understand perceptual demonstratives: terms like 'that mountain' or 'that book', used to refer to currently perceived objects. We use these terms in expressing perceptual-demonstrative beliefs and intentions, as when I say, 'I'm going to climb that mountain', or 'I've just found a mistake in that book'. I want, in particular, to focus on the way in which I understand someone else's uses of demonstrative terms. That is, I want to focus on the characterisation of my knowledge of someone else's demonstrative beliefs and intentions.

It has for some time been recognised that demonstrative propositions have a pivotal place in our ordinary thinking. By observing objects, we get knowledge first of demonstrative propositions concerning those objects; observation is in the first place a way of verifying demonstrative propositions. And if you are to act intentionally on an object, by reaching for it or grasping it, for example, then usually you need to have and use a demonstrative way identifying that object. If I want to pick up the box that contains the money, for example, I can do it only when I know the demonstrative identity, 'that box is the one with the money'.

Since demonstrative propositions have a pivotal place in our ordinary thinking, a knowledge of someone else's demonstrative thoughts will have a pivotal place in an understanding of that person's thinking. Knowledge of the other person's demonstrative thoughts will be essential to an understanding of what the other person is finding out from observation, and what actions the other person will perform – which box the person will pick up, for

instance. But how are we to characterise our knowledge of someone else's demonstrative thoughts?

One component in an understanding of someone else's demonstrative thoughts will be knowledge of the causal roles of those states: what causes a demonstrative belief or intention, for example, and which other states or actions it can cause. It is only because you know the causal roles of the other person's demonstrative thoughts that you can use this knowledge in explaining and predicting their actions, for example.

So how do you have knowledge of the causal roles of someone else's demonstrative thoughts? For the Theory Theory, grasp of the causal role of a psychological property is a matter of having tacit knowledge of a theory articulating the causal role of the property. For the simulation theory, grasp of the causal role of a psychological state is owed to a capacity that you have to simulate possession of that state. You can, in pretence or play or imagination, take on psychological states which are not your own, and in imagination follow through on the consequences of being in that state. This thinking is 'off-line' in that its upshot is not permanent, and it is decoupled from action. It provides you with a grasp of causal structure which puts to work the fact that your own states have a particular causal structure, rather than any representation of that causal structure.

There is pressure on the theory-theorist to weaken the notion of 'tacit knowledge of a theory' so that it becomes difficult to distinguish his view from the simulation theory. Suppose we consider the question whether someone is representing the following aspect of causal role: a belief that if *p*, then *q*, and a desire that *q*, will result in a desire to bring about *p*. Suppose that on various particular occasions our subject is confronted with the question what the causal upshot will be of having a belief that if *p*, then *q*, and a desire that *q*. On these occasions, our subject takes it that the upshot will be a desire to bring about *p*. What does it take for our subject to have tacit knowledge of the general principle, underlying those particular judgements? One approach is to take it that this is a question about the causal basis of the subject's judgements about the particular cases. The simplest response is to say that the general principle is tacitly known by the subject if there is a single common cause for all of these particular judgements about causal upshot by the subject. The problem with this approach here is this. Suppose the subject himself has a tendency to move from a belief that if *p*, then *q*, and a desire that *q*, to a desire to bring about *p*. And suppose that this tendency of the subject is systematic. That is, there is a single underlying common cause for all the times the subject makes this move in his own thinking. Now suppose that the subject makes judgements

about causal role by running simulations. Then whenever he runs a simulation of having a belief that if *p*, then *q*, and a desire that *q*, he concludes that this will cause a desire to bring about *p*. And by hypothesis, there is a single common cause for him reaching that result on all these occasions. So by our straightforward definition of 'tacit knowledge', this subject, by running simulations, counts as having tacit knowledge of causal role. The issue between the simulation theory and the Theory Theory has been lost.

Nevertheless, I do not want to pause too long over how we are to separate the Theory Theory and the simulation theory. I want rather to point out that there is a further dimension to knowledge of someone else's demonstrative thoughts. There is more to knowledge of someone else's demonstrative thoughts that knowledge of the causal roles of those thoughts. There is, in addition, a normative dimension. Given that someone's demonstrative thought actually has a given causal role, we can always ask whether it ought to have that causal role. We can ask whether these inputs really are enough for it to be right to form such a thought in response, and whether, given that you have such a thought, that really does make it right to draw such conclusions or execute such actions in consequence of having that thought. The Theory Theory and the simulation theory do not of themselves explain how there can be this normative dimension to your understanding of someone else's demonstrative thoughts, in addition to your knowledge of the causal roles of those thoughts. Of themselves, the Theory Theory and the simulation theory only offer views about the format in which you have knowledge of the causal roles of someone else's thoughts.

I am focusing specifically on the perceptual demonstratives in demonstrative thoughts, so the question particularly concerns the normative aspect of demonstratives. So the question is, how do we know what the right causal role is for a demonstrative to have? How do we know what ought to be demanded of the input to a thought containing the demonstrative? And how do we know what the implications of a thought containing the demonstrative ought to be?

The compelling answer is that we know these things by knowing the reference of the demonstrative. It is because you know which object the demonstrative stands for that you are in a position to assess the causal roles of thoughts containing the demonstrative.

But when you understand someone else's expression of a demonstrative thought, as when they say, 'that dog is on the loose again', how do you know which thing the person is referring to? The natural answer is that you do it by perceiving the thing. But it is not enough that the thing is in your field of view;

it might be in your field of view but quite unnoticed by you. You have to single it out visually. You have to attend to it.

Understanding someone else's use of a perceptual demonstrative does not, though, demand only that you and the other person should both be attending to the same object. After all it might be that you happen to be attending to the object anyway; that would not of itself mean that you knew what the other person was referring to. There is some sense in which you have to be jointly attending to the same object.

One line of thought here is that it should not be an accident that you are referring to the same object. The suggestion is that there should be some causal coordination between us, that results in it happening that we are attending to the same thing.

Another line of thought is that is should be transparent to us that we are attending to the same object; so far as the two of us are concerned, it should be out in the open, that we are attending to that object. The suggestion is that it should be common knowledge between us, which object we are attending to.

I shall discuss these to aspects of joint attention in turn, to try to determine which notion of joint attention we need in characterising knowledge of the normative aspect of demonstrative thoughts. I will argue that the notion of causal coordination is not directly relevant, but that the problematic notion of common knowledge is. Then I will return to the relation between this knowledge of the reference of a demonstrative thought, and the knowledge of causal role characterised by the Theory Theory and the simulation theory.

2. Causal coordination of attention

There are various notions of joint attention we can characterise that have to do with the way in which attention is controlled. The core of all these notions is that you and I are causally coordinating our attention onto the same object. It might be that we achieve attention to the same object through me slavishly attending to whatever you attend to, though you pay no heed to what I am attending to. Or it might happen that I attend to whatever I like and you track along behind me. But it can also be that our attention is reciprocally controlled: that you and I each keep track of what the other is attending to, so that we both work to ensure we attend to the same thing. As Tomasello puts it, 'In joint attention the child coordinates her attention to the object and the adult at the same time that the adult coordinates her attention to the same object and the child' (1995: 107).

Putting the notion of causally coordinated attention like this may make it sound as though we will have joint attention to an object only in the cases in which the object itself is of no intrinsic interest, so that it would not draw your attention to it of itself; you are each attending to it only because someone else is attending to it. But, of course, there can be more than one causal factor behind your attention to the object; it can be both partly stimulus-driven and partly driven by appreciation of what the other person is attending to. The important point, for the notion of causally coordinated attention, is that what the other person is attending to should be one of the causal factors making you attend to that thing.

So far, causal coordination of attention could be a fairly low-level matter, involving only sub-personal mechanisms so that both parties keep focused on the same thing. There are various ways in which propositional states could be involved in coordination. Propositional states might enter into the control of attention itself, and they might enter into my recognition of how my attention, or your attention, is being controlled. First, it might be that I know what you are attending to, and that this knowledge is a factor in sustaining my attention on the thing. Secondly, I might intend to attend to whatever you are attending to. And thirdly, it might be that I know that the reason I am attending to the thing is, in part, that you are attending to it. And finally, it might be that I know that the reason you are attending to the thing is, in part, that I am attending to it. With these materials in mind it would be possible to define a range of notions of joint attention which involved the psychological coordination of attention.

It should be borne in mind, though, that it is not just a theoretical possibility that there should be causal coordination of attention using only fairly low-level mechanisms. I was once running through a field and found a large group of cows gathered around a food trough in the corner of the field through which I had to go. I slowed to a walk, to assess the situation, and the cows simultaneously caught sight of me. They moved towards me, picking up speed, glancing at each other as they did so. They were engaging in social referencing, checking each others reactions to the object to which they were jointly attending. This joint attention did not require any of them to be engaging in conceptual thought; it was, presumably, a relatively low-level matter. Or again, a team playing football are continuously monitoring one another's attention. But this does not require them to be engaging in conceptual thought, or to have even iterated knowledge of the direction of each other's attention. Similarly, imitative learning may involve joint attention, but it does not seem to require conceptual thought or knowledge of what the other is attending to. If I learn how to

replace a flat tyre by watching you perform the operation, I may do it well as a result, but grasping the details may be an entirely non-conceptual matter, and so not involve knowledge of the direction of your attention at all.

It does not seem to me that any of these notions of joint attention as causal coordination, engaging though they are, are what we need in giving an analysis of what it is to understand another person's demonstrative thoughts. There is, as I said, another aspect to joint attention, which does not have to do with the control of attention, but rather with each person monitoring what the other person is attending to, so that there is common knowledge that they are attending to the same thing. To go back to an example Tomasello gives, suppose that you and I look out of our windows at a dog barking in the moonlight. We are attending to the same thing, but this does not count as joint attention because we may not even be aware of each other's existence. But suppose, as we lean out of our windows, that you and I catch sight of each other, so that there is common knowledge between us that we are both attending to the dog. This common knowledge that we are attending to the same thing has nothing to do with the control of attention. Perhaps you and I have quarrelled, so that the mere fact that you are looking at the dog has not the slightest tendency to dispose me to continue looking at it. The fact that you are looking at the dog may actually dispose me somewhat to look away, but in fact I remain fascinated by the commotion. So it is not that you are causing me to attend to the dog, or that I am causing you to. There is no 'coordination' between us. I would be attending to the thing whether or not you existed. There is only the mutual knowledge that we happen to be attending to the same thing. On the face of it, it is this notion of 'joint attention', depending not on control but on common knowledge, that matters for communication. If, as we look out of our windows, you say to me, 'that dog is barking', I will have no trouble in interpreting your remark, which thing you are talking about or what aspect of it you are commenting on, even though there is no coordination of the control of attention. And your remark might be intended to explain to me the meaning of the word 'barking' and so long as there is mutual knowledge of what we are attending to, the explanation will succeed.

Of course, the case of the barking dog is a bit unusual, in that the dog here is sufficiently striking to draw your and my attention independently. Often enough, all that brings it about that we are attending to the same thing is that we should be coordinating our attention, and it will be because we know that we are coordinating our attention that we know we are attending to the same thing. So it is not that coordination does not matter, but rather that it matters

only because it is instrumental for there being common knowledge between us that we are attending to the same thing.

3. The role of consciousness

I said that joint attention to the object is needed if you are to understand a perceptual demonstrative being used by someone else. It is striking that joint attention to an object generally involves both parties being conscious of the object. William James famously defined attention as follows:

> It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalisation, concentration of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others...
>
> (1890, vol. 1: 403–404)

On this commonsense understanding of the notion, attention is a phenomenon of consciousness. It is a modification of the conscious life: your experience is different when you attend to one object from when you attend to another object. Joint attention, then, we might think of as a further special modification of consciousness.

Suppose you think about the subject in a blindsight experiment, being asked to report on, say, the orientation of an object in the blind field. The experimenter can see the object perfectly well; the experimenter is consciously attending to it. There is also a sense in which the subject is attending to the object, even without being conscious of it. The subject is selectively reporting on the orientation of just that object, and if he reaches for the object or points to it he is selecting information from just that object to set the parameters for his actions. So even though he is not consciously attending to the object, there is an attenuated sense in which he is attending to it. There is even some causal coordination between the experimenter and the patient in their directing attention onto that object. Still, just because he is not conscious of the object, there is a way of understanding the experimenter's uses of perceptual demonstratives which is not available to the patient, when the experimenter says, for example, 'Can you point to that light'. The patient can construct descriptions which he can use to interpret the demonstratives the experimenter uses, descriptions like 'the light at which the experimenter is looking', or 'the light the experimenter wants me to talk about', so it is not as if their meaning is entirely opaque to him. But the ordinary way of understanding someone else's use of a perceptual

demonstrative does not require you to construct such a description. Ordinarily, you interpret someone else's use of a perceptual demonstrative just by being conscious of the object yourself. This point was already made by G. E. Moore when he wrote, 'Can we say "that thing"="the thing at which I am pointing" or "the thing to which this finger points" or "the nearest thing to which this finger points"? No, because the proposition is not understood unless the thing in question is *seen*' (*Commonplace Book* 1919–1953: 158). The joint attention we need to interpret each other's perceptual demonstratives involves conscious attention to the object.

I said that the reason we need joint attention to the object in order to have knowledge of each other's demonstrative thoughts is that joint attention is what supplies us with grasp of the normative dimension of the thought. It is because there is joint attention to the object that you can assess the causal role that the other person assigns to the thought. Joint attention is what supplies your knowledge of the reference of the thought. And it is because you know the reference of the thought that you can assess the evidence that led the other person to have that thought, and the actions or further reasoning which the other person will be engaged in as a consequence of having that thought. And you have knowledge of the causal role of the thought as a result either of your capacity to simulate the other person, or your having tacit knowledge of a theory about the causal roles that such thoughts have.

There is no particular reason to think that your knowledge of the causal role of someone else's demonstrative thought is itself conscious. It is true that simulation is often thought of as a way of generating verbal reports or conscious expectations about how someone will behave or how they may have come to their current condition. But this need not be the main or most primitive use of the capacity for simulation. In a recent paper, John Greenwood suggests that 'the characteristic output of simulation may be nonverbal (and perhaps non-conscious) anticipation (or expectation) of behaviour, rather than the verbal prediction of behaviour' (1999: 48). The suggestion is that this use of the capacity for simulation, in anticipating and responding to other people, may be what the system evolved to do, and may be more primitive and indeed more accurate than the linkage of the system with our conscious life and capacity for verbal report. Similarly, someone who thinks we have tacit knowledge of a theory of mind need not think that the theory has as its most primitive use the verbal reporting or the generating of conscious beliefs about the causal roles of other people's thoughts.

In a general way, we can distinguish between the causal relations that hold among the psychological states of a single person, and the causal relations that

hold between the psychological states of different people. Ordinary social interaction requires you to grasp both. Ordinary social interaction requires you to grasp facts about the dynamics of psychological states within a single person, such as the idea that one person's desire that *q*, and belief that if *p*, then *q*, may lead that person to desire that *p*. And ordinary social interaction requires you to grasp that if two people want an apple, and there is only one apple, that may lead to conflict.

The simulation theory and the Theory Theory, as usually formulated, relate only to grasp of the causal relations among the psychological states of a single person. Simulation will tell you about the dynamics of the psychological states only within the person being simulated, not about the causal relations holding between the psychological states of different people. Similarly, the Theory Theory, as usually formulated, will say something about the dynamics of the psychological states of a single person, but not about the causal relations between the psychological states of different people.

The natural way to develop Greenwood's suggestion is to say that the ability to simulate someone else has its most primitive role as input into the individual's capacity for social interaction: for collaborative activity, resolution of conflict and so on. This does not require that the use of simulation should itself be a conscious activity; it may be use, itself unconscious, of a system which feeds into the practical skills of social interaction. Similarly, tacit knowledge of a theory of mind might be thought to have its most primitive use in helping to guide the practical skills involved in social interaction. Of course, it may later be possible to make this knowledge available for conscious report, though we may be fallible in our attempts to make this kind of knowledge explicit.

The role of conscious joint attention will be to secure it that we have conscious access to the objectives of our interactive tasks, whether they are collaborative or competitive. This will mean that we have the basic knowledge of objectives which is required if we are to be capable of conscious criticism or assessment of the processes we are using to reach these objectives; knowledge of the processes being provided by simulation or theory.

4. Common knowledge

I want finally to look at epistemological issues around the place of common knowledge in an analysis of joint attention. The classical analysis of common knowledge, given by Lewis in *Convention*, is an infinite conjunction: a fact is common knowledge between you and me if we both know it, we both know

that we both know it, we both know that we both know that we both know it, and so on. For example, if you light a candle between us on a table, we may both be attending to the candle, and we may both know that we are, know that we know that we are, and so on.

In an ordinary case, how do I know what you are attending to? There are postural factors which might help; there is the characteristic hunched look of someone attending closely. If I see you crouched over a plant, inspecting it closely, that gives good reason to think that you are attending to that plant – I know which object is in question – even though that does not of itself answer all questions about the nature of your attention to the thing. There are also what we might call 'scenic' factors, factors relating to the nature of the scene before us. If you and I are sitting side by side before a screen, not looking at one another, and a single rapidly moving dot of light appears on the screen, then the thing is so salient that I may have a reasonable expectation that you are attending to it. So I don't have any postural evidence in this case, but the scene alone lets me generate a reasonable belief about what you are attending to. In this case I am implicitly appealing to the fact that we have a basic human similarity in what will pop out for us in this context, and that the moving dot of light would be found uniquely salient in that context by almost anyone. However, suppose a more complex scene appears on the screen. If I know that you are a big fan of Orson Welles, I may reasonably guess that your attention will have been riveted by Welles, though that isn't how it would go for absolutely everyone. And my specialist knowledge of you in particular may also play a role even in defining what range is of relevant objects in the scene, which constitute the options to which you might be attending. For example, if you are an academic just finishing a book on the history of hats, I might know that your first reaction will be to scan the hats, before settling on scrutiny of one, even though for most people the hats themselves do not define the objects to which they are likely to attend specifically.

Suppose that we have an optimum case, in which both the postural and scenic factors uniquely dictate the object. Say, you and I are sitting at a table with a candle between us which suddenly, unexpectedly, flares up a bit. It is so salient that I would expect it to draw your attention and your posture, leaping back while keeping your gaze fixed on the thing, confirms that you are attending to it. So I know that you are attending to the candle. Since you know my cognitive powers in this regard, you know that I will have realised that you are attending to the candle. Since I know that you appreciate my cognitive capacities, I know that you know I know you are attending to the candle. And so on. So we have the epistemic basis here for full infinitary common knowledge in

Lewis's sense. And of course, you can similarly see that I am attending to the candle, and I know you can see this, and so on. So we have common knowledge of what we are jointly attending to.

How much in the way of common knowledge is really needed for joint attention to give knowledge of the reference of a demonstrative? Suppose you say: 'that candle startled me'. How do I know which thought you have expressed? Is it not enough if I simply know what you are attending to? Why should it help to go through all these iterations?

In thinking about this topic it is quite confusing – or at any rate, I find it quite confusing – to stick with the everyday cases of demonstrative reference, because these are classical cases in which full common knowledge does seem to be available. I think it helps to consider somewhat unusual cases of demonstrative reference, in which we can control just how much knowledge can be had.

If I am to know which demonstrative thought you are having, there has to be some way in which it is indicated to me which thing you are thinking about. If you are expressing your thought to me, using a demonstrative term, there has to be what Kaplan called the demonstration accompanying the demonstrative – a pointing gesture, some descriptions indicating where to look, maybe a sortal term, and so on. The point about ordinary demonstrations is that they yield classical common knowledge scenarios. So suppose we consider a case in which the demonstration is a little bit unusual.

Suppose that you and I are both sitting in front of the same big screen, but in separate compartments, so that we cannot see or hear each other directly, though we can type messages onto the screen. And the screen contains a number of differently coloured circles. You type onto the screen, 'that circle is the same colour as Bill's hat'. How do I know which circle you are referring to? To let me know which circle you are attending to, you have a set of buttons which will light up lamps in my compartment. To signal that you are attending to the green circle, you push the green button and the green lamp lights up in my compartment. Then I know that you are attending to the green circle. The only problem is that the electronic apparatus only functions about 50% of the time; the signal only gets through about 50% of the time. So although I know that you are attending to the green circle, you do not know that I know that, since you do not know whether your message got through.

What I need, to let you know that your message got through, is an acknowledgement button, so I can light a lamp in your compartment confirming receipt of your message. Once you get that, you know that I know you are attending to the green circle. But again, the electronic apparatus only works

about 50% of the time, so I do not know whether you got my acknowledgement. You have to be able to confirm receipt of my acknowledgement; then I will know that you know I know you are attending to the green circle. And again, the apparatus only works about 50% of the time. And so on. Just to be fully explicit about this, the electronic apparatus never sends an incorrect message; it either works or sends no message at all. And you and I have no interest in sending wrong messages. So any message we get is accurate.

Suppose now that you type up on the screen, 'that circle is the same colour as Bill's hat', and push the green button to indicate to me that you are attending to the green circle. I do not send any acknowledgement. So now I know that you are attending to the green circle, and I am attending to it myself. But you do not know that I know what you are attending to. The iterations stop here. Is that enough for me to understand your use of the demonstrative, do I know which demonstrative thought you have expressed?

I think that we want some distinctions here. We can say that a full understanding of your thought would involve a full grasp of the causal role of the thought, and that common knowledge may be needed for there to be a full grasp of the causal role of the thought. But even at level 1 knowledge of what you are attending to, it is not as if I have no knowledge of the causal role of your thought; I may be able to simulate your condition exactly. What I miss has to do with the way in which your thought can be involved in social collaboration.

Suppose you and I are playing a war game, and we have to coordinate our attack on a single target. If we do manage to agree on the target, then victory is certain, but it will be a disaster if only one of us attacks; anything is better than that scenario. You type up on the screen, 'Let's attack that one at noon', pushing the green button to demonstrate to you which one you mean. The green lamp goes on in my compartment, so I know which one you have in mind. Does this make it rational for me to attack at noon? The problem is that you do not know whether your message has got through. So you do not know whether I will attack, so it is not rational for you to take the chance of attacking alone. So I do not know that you will attack, which makes it irrational for me to attack. I can try to improve the situation by pushing my acknowledgement button, but I will not know whether the message has got through, and you will know that I don't know that, so we will still be sitting tight. And so on; no finite iteration will allow us to manage the coordinated attack that is available to us and will secure victory. This kind of game is familiar from the literature on common knowledge. What its application to the case of joint attention and knowledge of someone else's demonstrative thoughts shows is this. The rational implications of knowledge of someone else's demonstrative thought do not depend only on

the content of the thought; they depend also on the extent to which it is out in the open that you do have knowledge of the content of the thought.

References

Greenwood, J. D. (1999). Simulation, Theory–Theory and Cognitive Peretration: No 'Instance of the Fingerpost'. *Mind and Language, 14* (1), 32–56.

James, W. (1890). *The Principles of Psychology*. New York: Holt.

Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Moore, G. E. (1962). *Commonplace Book 1919–1953*. London: Allen & Uwin.

Tomasello, M. (1995). Joint Attention as Social Cognition. In C. Moore & P. J. Dunham (Eds.), *Joint Attention*. Hillsdale, NJ: Erlbaum.

Reply to John Campbell

Elisabeth Pacherie
Institut Jean-Nicod, CNRS, Paris

John Campbell's objective in this paper is to examine what an understanding of someone else's perceptual demonstratives consists in and to offer a characterization of the cognitive capacities involved in such an understanding. He distinguishes between a causal dimension of our understanding of someone else's demonstrative thoughts and a normative dimension of this understanding. Along the causal dimension, understanding demonstrative thoughts involves knowing the causal roles of those thoughts, what causes them and what they can cause. How we come to such knowledge could be explained, according to Campbell, either by Theory Theory or by Simulation Theory. In the paper, he does not try to adjudicate between the two views. Actually, what he is really interested in is the normative dimension of our understanding of demonstrative thoughts. Here the question is not what the causal role of a given demonstrative thought is, but rather, given that the thought has a certain causal role, whether it ought to have it. Understanding a demonstrative thought along this normative dimension means being in a position to assess whether the causal role a demonstrative thought has is the right one for this thought to have. In order to be in such a position one must know which object the demonstrative stands for. Campbell claims that what is needed to gain such knowledge is joint attention to the object the demonstrative stands for. He then goes on to discuss which aspect of joint attention is central to an understanding of the normative dimension of the use of demonstrative thoughts. Is it causal coordination that is crucial or is it the common knowledge that we are attending to the same thing?

My own comments will also focus on the normative dimension of an understanding of demonstrative thoughts. I think Campbell's paper deals not just with one but with at least two normative issues. So I will try to separate them out. One question I will then consider is whether both issues are really specific

to demonstrative thoughts or whether they (or at least one of them) arise in communicative settings in general. Finally, once these distinctions are drawn, one may want to reconsider the question what common knowledge is really needed for.

Suppose I am standing in the middle of a room, surveying the furniture, and I exclaim: 'This armchair is really ugly!' There are two different kinds of situations in which this could happen. I may simply be expressing a thought, thinking aloud so to speak, while being quite oblivious of your presence, or I may be addressing you, trying to communicate a thought. It is obvious that it is the second kind of situation that Campbell has in mind. Yet drawing this contrast may help us distinguish among (1) aspects of the understanding of demonstrative thoughts that are common to the two kinds of situations, (2) aspects that are specific to the understanding of demonstratives in communicative settings, and, finally, (3) aspects of the understanding of thoughts, whether demonstrative or not that, that are specific to communicative settings. I agree with Campbell that our knowledge of demonstrative thoughts involves both causal and normative components, but we may have to distinguish more than one causal role and also more than one normative aspect.

Suppose we are in the first situation: when saying that this armchair is ugly, I am simply thinking aloud, oblivious of your presence; I may be expressing a thought but there is no communicative intent involved. In such a case, a causal understanding of my demonstrative thought will involve finding out what the causal roles of beliefs about the ugliness of armchairs are. In turn, a normative understanding of this thought will involve finding out what object I am referring to and whether it is indeed an ugly armchair.

In his discussion of joint attention, Campbell distinguishes two desiderata. First, it should not be an accident that you, the interpreter, and I, the utterer, are referring to the same object, so there should be some coordination between us. Second, it should be transparent to us that we are attending to the same object, so there should be common knowledge between us regarding which object we are attending to. Campbell further claims that the first requirement is subordinate to the second. Coordination matters only because it is instrumental for there to be common knowledge between us that we are attending to the same thing.

Clearly, this analysis does not apply to the situation at hand. When exclaiming about the ugliness of the armchair, I am not trying to communicate anything to you. I may not even be aware of your presence. Yet, in favourable circumstances, you will be able to come to an understanding, both causal and normative, of my demonstrative thoughts. This understanding will not require

the transparency condition to obtain, but will rest entirely on some form of causal coordination of attention. So it seems that you can have some normative understanding of my demonstrative thoughts without any knowledge on my part that you have this understanding. Note, however, the qualification 'in favourable circumstances'. Since when exclaiming about the ugliness of the armchair I am not trying to communicate anything, my utterance may not be accompanied by any overt indication of what I am attending to. For instance, if there are several armchairs in the room, it might not be clear which one of them I have in mind.

This is one important aspect of the contrast with the second situation. Suppose now that my remark about the armchair was made with a communicative intent. In order to understand it, there will be two causal roles for you to elucidate: the causal role of the thought I am communicating and the causal role of my intention to communicate this thought. Now the presence of an intention to communicate makes a big difference. Your understanding of the normative dimension of my demonstrative thought will rest in part on your understanding of the causal role of my intention to communicate this thought. For this communicative intention will lead me to try to make sure that you understand which object the demonstrative I am using stands for. Understanding that my intention has that causal role allows you to interpret with confidence the overt signs that I am attending to a given object as reliable indications that this is the object my demonstrative stands for. I would like to make three remarks about this second type of situation.

According to Campbell, either the Simulation Theory (ST) or the Theory Theory (TT) could explain how you could have knowledge of the causal role of my thoughts. John also points out, later in the paper, that the Simulation Theory and the Theory Theory, as usually formulated, relate only to grasp of the causal relations among the psychological states of a single person. I just mentioned that in a communicative setting there are two causal roles to be elucidated. The causal role of the thought I am communicating and the causal role of my intention to communicate this thought. My first remark is about the causal role of communicative intentions. This causal role involves causal relations that hold between the psychological states of at least two people. According to the classical Gricean analysis of communicative intentions, I intend, among other things, that you recognize my intention to get you to think that p.[1] So if either ST or TT is to explain how one can understand the causal role of communicative intentions, they will have to tell us about the dynamics of the causal relations holding among the psychological states of different people. Campbell suggests in his paper that Greenwood's ideas about simulation could

be developed in order to account for our capacity for social interaction. There is another line of thought that I would like to briefly mention. We use simulation not just to understand and predict the behaviour of other people but also when planning our own actions. As Michael Bratman[2] has remarked, even in the case of individual actions intrapersonal coordination is needed. Individual intentions and actions are subject to consistency constraints. Plans must be internally consistent: the intentions that are the constitutive elements of a plan must be mutually consistent, for it to be possible that the entire plan be successfully executed. Plans must also be 'externally consistent', that is consistent with the beliefs of the agent about the world she is in. Conflicts of intentions do not arise solely in social contexts. A single agent may well have conflicting intentions and he needs procedures to resolve these conflicts. My point here is simply that the dynamics of the psychological states within a single agent is already quite complex and that there might not be an enormously wide gap between grasping the causal relations that hold among the states of a single agent and grasping the relations that hold among the states of different agents.

My second remark about the understanding of demonstrative thoughts in communication concerns the status of the transparency condition insisted on by Campbell. Is transparency, hence mutual knowledge, really necessary for a normative understanding of someone else's demonstrative thoughts? Certainly mutual knowledge appears to be sufficient for the interpreter to grasp the reference of the demonstrative hence to have a normative understanding of the thought. But mutual transparency is perhaps not required. For the interpreter to grasp the normative dimension of the utterer's demonstrative thought, it must indeed be clear to him that he is attending to the same object that the utterer is attending to, but the reverse does not seem to hold. Suppose when I exclaimed "this armchair is really ugly", I was visiting a castle and trying to engage in conversation with you, a fellow visitor. Since this castle attracts a lot of tourists, I am not even sure that you speak English. Let us suppose you do indeed speak English but don't want to engage in a conversation with some unknown and probably boring tourist, so you don't react. The mutual knowledge condition is not satisfied, since I have no clue whether or not you are attending to the armchair I am commenting about. Yet, this is no obstacle to your understanding of the normative dimension of my demonstrative thought. So, to summarize, it seems that:

1. One-way transparency is really all that is needed for normative understanding.

2. One-way transparency is a consequence of the utterer's intention to communicate: this communicative intention will lead the utterer to try to make it transparent to the addressee which object the demonstrative he is using stands for.
3. The interpreter's ability to exploit this one-way transparency rests in part on his understanding of the causal role of communicative intentions.

My third and last remark concerns mutual knowledge: what is mutual knowledge needed for? I think mutual knowledge is not required for an understanding of demonstrative thoughts per se. Rather it is required for wider collaborative endeavours in which the communication of demonstrative thoughts is used as a means for some further ends. There are indeed cases where not only successful communication, but also knowledge that communication is successful are pre-conditions of the success of some larger collaborative activities of which the communicated thoughts are part. So if communication involves demonstrative thoughts, indeed the success of communication implies that the interpreter should have grasped which object the demonstrative stands for and knowledge of the success of communication implies mutual knowledge on the part of both the utterer and the interpreter that they are attending to the same thing.

The war game discussed by Campbell at the end of his paper is indeed a prime instance of such a situation. But we may wonder whether demonstrative thoughts are not just a special case of a much more general problem. In the case of demonstrative utterances, the message I want to convey consists of a string of words together with some demonstration. The way Campbell sets the situation is such that there are two channels of communication: computer screens used for the communication of strings of words and lights used both for demonstrations and acknowledgements. Only the second channel is unreliable.

But now take a standard version of the war game,[3] with only one communication channel, also unreliable. The same type of coordination problem will arise and it will arise even if the players are using non-demonstrative utterances (say, if they use a proper name, 'hill 135' or a definite description, 'the circle on the upper left corner' to refer to the target). So, indeed the use of demonstrative utterances adds a new twist – we need reliable channels both for the strictly linguistic material and for the demonstrations – but the problem of mutual knowledge is a general problem whose scope is not restricted to the understanding of demonstrative thoughts.

John Campbell concludes his paper with the following sentence: "The rational implications of knowledge of someone else's demonstrative thought do

not depend only on the content of the thought; they depend also on the extent to which it is out in the open that you do have knowledge of the content of the thought". What I am suggesting is that we should distinguish the normative dimension involved in the understanding of demonstrative thoughts per se from the normative dimension having to do with the use of demonstrative thoughts in an attempt at some collaborative endeavour. My own conclusion would therefore be that we should distinguish between the rational implications of knowledge of someone else's demonstrative thought and the rational implications of knowledge of the intention with which someone else communicates a demonstrative thought. It is the latter that depends on the extent to which it is out in the open that you do have knowledge of the content of the thought.

Notes

1. Actually, there are counter-examples to the Gricean analysis suggesting that more iterations are needed. But we may leave this complication aside here.

2. Bratman, M. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

3. The wargame is an unusual type of collaborative endeavour and this in two different ways: first because of the peculiar conditions of communication, second because it is a one-step game, with the consequence that coordination must be achieved right from the start. The second feature makes initial mutual knowledge a necessary condition of success and the peculiar conditions of communication make mutual knowledge impossible. Most collaborative ventures are different: several steps are involved and this allows for the progressive establishment of coordination and mutual knowledge.

Subject index

Name index

In the series ADVANCES IN CONSCIOUSNESS RESEARCH (AiCR) the following titles have been published thus far or are scheduled for publication:

1. GLOBUS, Gordon G.: *The Postmodern Brain.* 1995.
2. ELLIS, Ralph D.: *Questioning Consciousness. The interplay of imagery, cognition, and emotion in the human brain.* 1995.
3. JIBU, Mari and Kunio YASUE: *Quantum Brain Dynamics and Consciousness. An introduction.* 1995.
4. HARDCASTLE, Valerie Gray: *Locating Consciousness.* 1995.
5. STUBENBERG, Leopold: *Consciousness and Qualia.* 1998.
6. GENNARO, Rocco J.: *Consciousness and Self-Consciousness. A defense of the higher-order thought theory of consciousness.* 1996.
7. MAC CORMAC, Earl and Maxim I. STAMENOV (eds): *Fractals of Brain, Fractals of Mind. In search of a symmetry bond.* 1996.
8. GROSSENBACHER, Peter G. (ed.): *Finding Consciousness in the Brain. A neurocognitive approach.* 2001.
9. Ó NUALLÁIN, Seán, Paul MC KEVITT and Eoghan MAC AOGÁIN (eds): *Two Sciences of Mind. Readings in cognitive science and consciousness.* 1997.
10. NEWTON, Natika: *Foundations of Understanding.* 1996.
11. PYLKKÖ, Pauli: *The Aconceptual Mind. Heideggerian themes in holistic naturalism.* 1998.
12. STAMENOV, Maxim I. (ed.): *Language Structure, Discourse and the Access to Consciousness.* 1997.
13. VELMANS, Max (ed.): *Investigating Phenomenal Consciousness. Methodologies and Maps.* 2000.
14. SHEETS-JOHNSTONE, Maxine: *The Primacy of Movement.* 1999.
15. CHALLIS, Bradford H. and Boris M. VELICHKOVSKY (eds.): *Stratification in Cognition and Consciousness.* 1999.
16. ELLIS, Ralph D. and Natika NEWTON (eds.): *The Caldron of Consciousness. Motivation, affect and self-organization – An anthology.* 2000.
17. HUTTO, Daniel D.: *The Presence of Mind.* 1999.
18. PALMER, Gary B. and Debra J. OCCHI (eds.): *Languages of Sentiment. Cultural constructions of emotional substrates.* 1999.
19. DAUTENHAHN, Kerstin (ed.): *Human Cognition and Social Agent Technology.* 2000.
20. KUNZENDORF, Robert G. and Benjamin WALLACE (eds.): *Individual Differences in Conscious Experience.* 2000.
21. HUTTO, Daniel D.: *Beyond Physicalism.* 2000.
22. ROSSETTI, Yves and Antti REVONSUO (eds.): *Beyond Dissociation. Interaction between dissociated implicit and explicit processing.* 2000.
23. ZAHAVI, Dan (ed.): *Exploring the Self. Philosophical and psychopathological perspectives on self-experience.* 2000.
24. ROVEE-COLLIER, Carolyn, Harlene HAYNE and Michael COLOMBO: *The Development of Implicit and Explicit Memory.* 2000.
25. BACHMANN, Talis: *Microgenetic Approach to the Conscious Mind.* 2000.
26. Ó NUALLÁIN, Seán (ed.): *Spatial Cognition. Selected papers from Mind III, Annual Conference of the Cognitive Science Society of Ireland, 1998.* 2000.
27. McMILLAN, John and Grant R. GILLETT: *Consciousness and Intentionality.* 2001.

28. ZACHAR, Peter: *Psychological Concepts and Biological Psychiatry. A philosophical analysis.* 2000.
29. VAN LOOCKE, Philip (ed.): *The Physical Nature of Consciousness.* 2001.
30. BROOK, Andrew and Richard C. DeVIDI (eds.): *Self-reference and Self-awareness.* 2001.
31. RAKOVER, Sam S. and Baruch CAHLON: *Face Recognition. Cognitive and computational processes.* 2001.
32. VITIELLO, Giuseppe: *My Double Unveiled. The dissipative quantum model of the brain.* 2001.
33. YASUE, Kunio, Mari JIBU and Tarcisio DELLA SENTA (eds.): *No Matter, Never Mind. Proceedings of Toward a Science of Consciousness: Fundamental Approaches, Tokyo, 1999.* 2002.
34. FETZER, James H.(ed.): *Consciousness Evolving.* 2002.
35. Mc KEVITT, Paul, Seán Ó NUALLÁIN and Conn MULVIHILL (eds.): *Language, Vision, and Music. Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, 1999.* 2002.
36. PERRY, Elaine, Heather ASHTON and Allan YOUNG (eds.): *Neurochemistry of Consciousness. Neurotransmitters in mind.* 2002.
37. PYLKKÄNEN, Paavo and Tere VADÉN (eds.): *Dimensions of Conscious Experience.* 2001.
38. SALZARULO, Piero and Gianluca FICCA (eds.): *Awakening and Sleep-Wake Cycle Across Development.* 2002.
39. BARTSCH, Renate: *Consciousness Emerging. The dynamics of perception, imagination, action, memory, thought, and language.* 2002.
40. MANDLER, George: *Consciousness Recovered. Psychological functions and origins of conscious thought.* 2002.
41. ALBERTAZZI, Liliana (ed.): *Unfolding Perceptual Continua.* 2002.
42. STAMENOV, Maxim I. and Vittorio GALLESE (eds.): *Mirror Neurons and the Evolution of Brain and Language.* n.y.p.
43. DEPRAZ, Natalie, Francisco VARELA and Pierre VERMERSCH.: *On Becoming Aware.* n.y.p.
44. MOORE, Simon and Mike OAKSFORD (eds.): *Emotional Cognition. From brain to behaviour.* 2002.
45. DOKIC, Jerome and Joelle PROUST: *Simulation and Knowledge of Action.* 2002.
46. MATHEAS, Michael and Phoebe SENGERS (ed.): *Narrative Intelligence.* n.y.p.
47. COOK, Norman D.: *Tone of Voice and Mind. The connections between intonation, emotion, cognition and consciousness.* 2002.
48. JIMÉNEZ, Luis: *Attention and Implicit Learning.* n.y.p.
49. OSAKA, Naoyuki (ed.): *Neural Basis of Consciousness.* n.y.p.