1 Introduction

Peter Carruthers and Peter K. Smith

Since Premack and Woodruff’s classic (1978) paper, the term ‘theory of mind’ has come to designate a particular research-domain, whose goal is to provide an explanation of the ability – which may or may not be unique to human beings – to explain and predict the actions, both of oneself, and of other intelligent agents. It is also the term used to designate one controversial characterisation of the basis of this ability. So-called ‘theory-theorists’ maintain that the ability to explain and predict behaviour is underpinned by a folk-psychological theory of the structure and functioning of the mind – where the theory in question may be innate and modularised, learned individually, or acquired through a process of enculturation. There are, then, many different theories of theory of mind in its research-domain sense; and the theory-theory account of our theory of mind abilities, too, admits of a number of different variants. Moreover, both sets of questions are now of common concern to a number of different disciplines, including philosophy, primatology, cognitive science, and developmental psychology. These multiplicities are reflected in the title of this volume, and in its contents. In this brief introductory chapter we shall provide the reader with a map of the area, laying out the issues and options, and linking these to the chapters that follow.

I A recent history

Curiously, the modern phase of work on theory of mind by developmental psychologists began with a paper in primatology – Premack and Woodruff, 1978 – which raised the question whether chimpanzees have a theory of mind. This question proved immensely difficult to answer, although it has spawned a good deal of very interesting research since (continued in Part IV of this volume). But what it certainly did do, was to force all psychologists to think very hard about what it is to possess a conception of the mind of another creature, and also about the behaviour which might show whether or not such a conception is possessed. This thinking bore fruit a few years later (seeded, it should be said, by suggestions from the
philosophers Daniel Dennett and Gilbert Harman in their commentaries on Premack and Woodruff’s paper) with the publication of Wimmer and Perner’s 1983 article, containing the now-famous false-belief task for use with children – about which more will be said shortly.

In fact the developmental study of theory-of-mind skills in children had had, in a sense, a history going back at least to the seminal work of Piaget on children’s thinking and egocentrism in the 1930s to 50s, and the following work on empathy and role-taking skills in the 60s and 70s. But while this research was of value, the topic did not ‘take off’ then in the way that it has in the 80s and 90s. There may be several reasons for this. First, until the late 1970s developmental psychologists in this tradition were perhaps too narrowly focused within the framework of Piagetian theory – whether aiming to confirm, or disprove, its tenets. Second, and relatedly, there was not then the influx of ideas from philosophy, primatology, and cognitive science which has proved so fruitful more recently. And third, the research lacked the focus of a clear experimental paradigm, which only appeared with the invention of the false-belief task in the early 80s.

The original false-belief task involved a character, Maxi, who places some chocolate in a particular location and then leaves the room; in his absence the chocolate is then moved to another location. The child is then asked where Maxi will look for the chocolate on his return. In order to succeed in this task, the child must understand that Maxi still thinks that the chocolate is where he left it – the child must understand that Maxi has a false belief, in fact. The task works, because in order to succeed in it the child must be able to contrast its own perception of the real situation with the belief of the target agent – in order to get the right prediction, the child must be capable (in one way or another), not just of representing the state of the world, but of representing Maxi’s representation of the world.

The surprising initial finding was that children became able to succeed in this task between the ages of 4 and 5, some two or three years earlier than Piaget would have predicted. A small industry then developed, replicating and refining the task, and attempting (with some success) to push the critical age of acquisition still lower. Other sources of evidence also began to be looked at – for example, relating to children’s early verbalisations concerning mental states, their understanding of emotions, and their ability to deceive others. And researchers began to investigate the developmental stages through which theory-of-mind abilities are acquired. (For extensive discussion, bold theorising, and reviews of the literature, see Wellman, 1990 and Perner, 1991a.) Others continued to investigate the possibility of theory-of-mind abilities in non-human primates. And yet others again began to look at the possible absence of theory of mind in some humans, particularly in connection with autism (see Baron-Cohen et al., 1985).
Most psychologists through this period have remained neutral on the nature of the end-state of theory of mind acquisition. Many, for example, have been careful to speak of theory of mind (or mind-reading) abilities, without commitment as to whether these abilities are underpinned by anything resembling a theory. But increasingly there has been a realisation that questions concerning the process of theory of mind acquisition cannot be kept wholly separate from questions concerning the nature of what is acquired, and a number of developmental psychologists have taken stands on the latter issue. For example, Wellman (1990) and Gopnik (1993) have declared themselves for a theory-theory account of our mature mind-reading abilities, and Harris (1989) has defended the simulationist alternative. (More will be said about the contrast between theory-theory and simulationism shortly.)

Philosophers, of course, have had a long-standing interest in the basis of our (adult) knowledge of the minds of ourselves and others, going back to Descartes and beyond. With the demise of Cartesianism in the 1940s and 50s, and the collapse of behaviourism in the 60s, some or other form of theory-theory has come to be dominant, at least since David Lewis’ classic (1966) paper. According to this view, our access to the minds of other people (and also our access to our own mental states) is mediated by an implicitly held theory of the structure and functioning of the human mind; and the different types of mental state are individuated by their functional role as described within this theoretical system. Until the mid-1980s philosophers were not overly interested in the process of acquisition, as such, assuming it to involve some or other form of enculturation. But that has now changed dramatically, for two distinct reasons.

The first is that philosophical interest in theory of mind has recently come to intersect with another perennial debate, between Rationalists and Empiricists, concerning innate versus learned knowledge; and this debate is, of course, at least partly a developmental one. In particular, Fodor (1981, 1987, 1992) has proposed that our knowledge of folk-psychological theory is innate, developing in the child through a process of maturation rather than of learning.

The second reason arises out of the challenge mounted to the theory-theory orthodoxy by Robert Gordon (1986) and, independently, by Jane Heal (1986), who each proposed a simulationist alternative. According to this view, what lies at the root of our mature mind-reading abilities is not any sort of theory, but rather an ability to project ourselves imaginatively into another person’s perspective, simulating their mental activity with our own. This view has been developed further by Gordon (1992a) and cultivated rather differently by Goldman (1989, 1992b, 1993a). And it rapidly began to dawn on people that simulationism and theory-theory might have
different implications for (and be tested against) the process of normal
theory of mind development; and also that they might provide competing
explanations of some of the experiences and difficulties of people with
autism.

By 1992 the time was then ripe for philosophers and psychologists to
begin to interact with one another much more directly and fruitfully on
these issues; and the stage was thus set for the project leading up to the
present volume. But before introducing the reader to its contents, it may be
helpful to lay out the issues and options in a little more detail.

2 A map of the area

We focus first upon the end-state of theory of mind acquisition, namely
how a mature adult explains and predicts the actions of self and others. The
basic theoretical choice is between theory-theory and simulationism, with
also the possibility of some form of simulation/theory mix. Each of these
options then admits of further sub-divisions. If mature theory of mind abil-
ities are underpinned by a theory, then this theory may be more or less
implicit (more or less accessible to consciousness and oral report); and it
may be more or less deeply theoretical (for example, consisting of a more
or less tightly structured and inter-related set of principles). If those abil-
ities are underpinned by simulation, on the other hand, then one can either
maintain that simulation presupposes first-person awareness of one's own
mental states, with the inference from self to other being a kind of argu-
ment from analogy (as Goldman has claimed); or one can maintain that
simulation involves a kind of imaginative identification which can operate
without introspective self-awareness (as Gordon has claimed). And then of
course there are a whole range of possible simulation/theory mixes, main-
taining that some specified components of our theory of mind abilities are
underpinned by simulation and others by theory.

A separate issue is how these theory of mind abilities develop; although
the theoretical choices above have implications for the process of develop-
ment. It is hard to see, for example, how anyone could consistently main-
tain that the end-state is one of simulation, while claiming that theory of
mind acquisition is a process of theorising. For how could theorising lead
to something which is not a theory, namely an ability to simulate? It is
almost as hard to see how an ability to simulate could be acquired entirely
through enculturation, or social learning. In fact, simulationists maintain
that the ability to simulate is grounded in an innate genetic endowment.
This is variously characterised as the ability to imagine; the ability to think
counter-factually; the ability to entertain suppositions; or the ability to take
one's practical reasoning system 'off-line'. But simulationists would also
allow that there is an element of learning involved in development, since children have to learn which of their mental states to vary when they simulate another person, in order to constitute their then-altered perspective.

It is almost equally hard to see how anyone could maintain that the end-state is one of theory while the process of acquisition is one of simulation. For why should simulation drop out once we have attained mature theory of mind abilities (by hypothesis, a mature theory), if it was nevertheless the crucial engine of acquisition? And, indeed, there is no one who attempts to defend this option. Those who are theory-theorists about the end-state mostly face a three-way choice concerning the mechanism of acquisition: it can either be a matter of growth in an innate module; it can be learning by theorising; or it can be learning through teaching and enculturation (or any pair-wise combination, or indeed any combination of all three processes).

Some theory-theorists maintain that the process of acquisition is one of biological growth. They believe that folk-psychology (or at least some of the core components thereof) is embodied in an innate, genetically endowed, theory of mind module which grows in the normally developing child (perhaps passing through a number of intermediate theory-like stages). Normal development may require triggering experiences from the environment, of course; and these may include the experience of enculturated talk about the mind. And there may also be an element of learning around the periphery. But on this view, the core process is not one of learning or theorising.

Other theory-theorists maintain that theory of mind is learned on the basis of experience. In one version, the child is pictured as a little scientist, constructing and revising theories in the light of incoming data. In the other version, the child is seen more as a little science student, picking up the folk-psychology of its culture through interacting with – and listening to the talk of – its carers and older siblings. Almost everyone now holding one or other of these positions maintains that there is some innate initial basis – some innate starting point of attention-biases or similarity-spaces which gives a kick-start to the process of theory of mind acquisition. But that process itself is conceived to be one of learning.

Finally, those who maintain some or other form of simulation/theory mix for the end-state of development are free to adopt almost any combination of the above possible processes of acquisition. No wonder that the theory of mind literature has grown increasingly complex, and that the issues are so difficult to resolve!

A widely accepted hypothesis is that autism involves at least some sort of theory of mind deficit. People with autism have notorious problems in communication and social interaction, and in reading the minds of other
people—well-documented in their difficulties with false-belief tasks, for example. And the various proposals canvassed above, concerning the process of normal theory of mind acquisition, carry different implications for the understanding of these phenomena. For simulationists, autism must result (at least in part) from a damaged capacity to imagine, or to engage in counter-factual reasoning. For modularist theory-theorists, autism will result (again, perhaps only in part) from a damaged or destroyed theory of mind module. And for either sort of learning theory-theorist, it must either be the capacity to learn from experience which is damaged, or autism may result from the lack of (or damage to) the innate initial basis of normal development. This is then fertile testing-ground for a whole range of specific hypotheses concerning normal theory of mind development.

The question whether or not non-human primates possess theory of mind abilities connects with a number of the above issues in a variety of complex ways. For example, if chimpanzees (or other great apes) were to turn out to possess some simpler version of our human folk-psychology, then this might contain valuable clues to the underlying basis for, and also the evolution of, our own capacities. But part of the interest of the primate research is also methodological. For any tests of theory of mind competence in non-human primates must, perforce, be non-language-involving; and the construction of such tests may then cast important new light on the course of theory of mind development in young children.

3 A guide to the volume

Parts I and II of the volume divide, very roughly, into chapters on the end-state of theory of mind development, and chapters focusing on the process of development itself; though often with very considerable overlap. Many of the chapters in Part I are distinctive in proposing some form of simulation/theory mix (see Carruthers, ch. 3; Heal, ch. 5; Perner, ch. 6; Botterill, ch. 7).

Gordon (ch. 2) presents and argues for a radical form of simulationism, according to which our very concepts of the mental are acquired through a process of simulation, without subjects needing to have introspective access to their own mental states as such—though he is now prepared to allow a limited place for introspection of what are in fact mental states, but not initially categorised as mental, but rather as states of the body. Carruthers (ch. 3) then criticises both the Gordon and Goldman versions of simulationism, on the grounds that neither can account adequately for our introspective knowledge of our own occurrent thought-processes, in particular. (Gordon, he alleges, is a quasi-behaviourist, whereas Goldman is claimed to be a sort of Cartesian.)
Nichols et al. (ch. 4) explore the explanatory potential of off-line simulation accounts of a variety of cognitive capacities, to compete with the sort of information-based explanations more usual in cognitive science. They consider how off-line simulation might be appealed to in explanation of counter-factual reasoning, of empathy, and of mental imagery, as well as considering its more usual role in purported explanations of our theory of mind abilities. In the latter connection, they present new evidence that behaviour prediction is not done by simulation; but they argue that the other three possibilities are well worth further exploration. Besides an explicit criticism of the simulationist account of our theory of mind abilities, this chapter also contains an implicit criticism. For many simulationists have argued for their position on grounds of simplicity, claiming that off-line simulation will need to be appealed to in any case, to explain counter-factual reasoning and empathy, for example. Nichols et al. attempt to undercut this argument, since they clearly think that it will be a different system which is taken off-line in each case.

Heal (ch. 5) and Perner (ch. 6) both argue for a simulation/theory mix. Heal argues that while simulation must be employed to enable us to predict what people will come to believe on the basis of their other beliefs, we must also possess a body of theoretical knowledge concerning the causal roles of various types of propositional attitude. Perner makes a new proposal to distinguish simulation from tacit theory, and presents a range of experimental data to suggest that simulation cannot be by any means the whole story.

Botterill (ch. 7) makes a useful distinction between different versions of theory-theory, depending upon how deeply theoretical the body of information is supposed to be, and argues tentatively in favour of strong theoreticity. Stone and Davies (ch. 8) then lay out their view of the State of the Art, meanwhile commenting on the other chapters in Part I.

Segal (ch. 9), and Baron-Cohen and Swettenham (ch. 10), open Part II of the volume by defending and developing modularist forms of theory-theory. Segal begins his chapter by distinguishing a number of different notions of ‘modularity’; he then contrasts one sort of modularist account of development with the child-as-scientist alternative, and argues on a variety of grounds that the former is preferable. Baron-Cohen and Swettenham assume the correctness of a modularist framework, and explore the relationship between the theory of mind module (ToMM) and a postulated precursor module, the shared attention mechanism (SAM), using the results of a recent large screening study in which young children were tested for autism.

Gopnik (ch. 11) and Astington (ch. 12) are both learning-theorists; but of different sorts. Gopnik takes issue with modularity accounts, and
defends her version of the child-as-scientist account of theory of mind acquisition. Astington argues that renewed consideration should be given to enculturation accounts – placing greater emphasis on learning from adults and on the role of language in theory of mind acquisition – and presents a range of empirical data in their support.

Harris (ch. 13) raises a little-noticed problem, which is a difficulty for simulationists and theory-theorists alike, namely: why is it that beliefs are so much more difficult for children to understand than desires? His novel proposal is that the notion of desire has its essential place in the understanding of agency, which comes more easily to children because they are acquainted with their own agency from the start; whereas the notion of belief has its locus in the understanding of communication, requiring some linguistic competence as a prerequisite.

The three chapters in Part III focus on autism. Boucher (ch. 14) presents and defends the criteria for an adequate explanation of autism, arguing that theory-theory explanations fall short of a number of them. Her view is that autism probably has multiple bases. Currie (ch. 15) compares and contrasts theory-theory with simulationist explanations of autistic phenomena, arguing tentatively for the superiority of the latter. And finally Carruthers (ch. 16) tries to show how a theory-theorist can explain features of the autistic syndrome (specifically, absence of pretend play and executive function deficits) which might otherwise appear problematic for the ‘mind-blindness’ theory of autism.

The four chapters in Part IV round out the volume by considering a variety of issues arising from the domain of primatology. Whiten (ch. 17) considers a number of proposals for marking the distinction between clever behaviourism and genuine mentalism. Povinelli (ch. 18) explores at length the prospects and potential significance of discovering theory of mind abilities in non-human primates, and presents dramatic new evidence suggesting that chimpanzees do not even have an understanding of visual attention. Gómez (ch. 19) discusses some of the difficulties inherent in testing for theory of mind abilities in non-human primates, and presents the results of experiments conducted with a captive orang-utan. Finally, Smith (ch. 20) concludes the volume with a comment on the previous chapters in Part IV, and with the suggestion (similar to the one made by Harris in ch. 13) that theory of mind abilities may be dependent upon language.

This volume may not have definitively resolved any of the outstanding issues concerning the nature of theory of mind abilities and their attainment. But it has at least clarified those issues, and put up a number of new possibilities worthy of further investigation. It also provides an exemplar, we believe, of the sort of inter-disciplinary collaboration necessary for future progress in this area.