



The theory of mind module in evolutionary psychology

PHILIP GERRANS

Department of Philosophy, Adelaide University, Adelaide, South Australia 5005, Australia (e-mail: philip.gerrans@adelaide.edu.au)

Received 4 September 2001; accepted in revised form 4 September 2001

Key words: Theory of Mind, Evolutionary Psychology, Autism, Modularity

Abstract. Evolutionary Psychology is based on the idea that the mind is a set of special purpose thinking devices or modules whose domain-specific structure is an adaptation to ancestral environments. The modular view of the mind is an uncontroversial description of the periphery of the mind, the input-output sensorimotor and affective subsystems. The novelty of EP is the claim that higher order cognitive processes also exhibit a modular structure. Autism is a primary case study here, interpreted as a developmental failure of a module devoted to social intelligence or Theory of Mind. In this article I reappraise the arguments for innate modularity of TOM and argue that they fail. TOM ability is a consequence of domain general development scaffolded by early, innately specified, sensorimotor abilities. The alleged Modularity of TOM results from interpreting the outcome of developmental failures characteristic of autism at too high a level of cognitive abstraction.

Introduction

Evolutionary Psychology (EP) explains human cognition by treating our psychological phenotype as an adaptation to ancestral environments. Crucial to this project is the idea that the human mind is not an all purpose cognitive device but a set of devices specialised for specific cognitive tasks. “Even a restricted consideration of hunter gatherer tasks suggests that it is unlikely that any single general computational system could solve them all under ancestral conditions. (Indeed it is difficult to imagine a domain general computational system that could solve any of them)” (Cosmides and Tooby 1994, p.90.) Vindication of the project of EP thus depends on the discovery within our psychological phenotype of such specialised cognitive mechanisms, or modules.

Everyone accepts that there is an interesting evolutionary story to be told about the origin of peripheral processing modules such as vision and face recognition. The boundary of the mind is highly structured and exploits specialised computational devices to detect and respond to perceptual stimuli. This specialisation, however, comes at a price. A mind composed entirely of such perceptual and quasi-perceptual input and output modules would be quite rigid, stimulus bound, and unable to cope with novel situations which exceed its online first-order representational capacities. Sperber (1994) gives the hypothetical example of a ground dwelling rodent with evolved systems specialised to detect the vibration caused by heavy-footed preda-

tors which, when activated, cause the animal to flee to its burrow. In a coyoteless contemporary world the same systems are maladaptive. The rodent starves in its burrow because of the constant rumble of trains across the prairie.

Humans, however, are not so stimulus bound. They can form beliefs about the veridicality of their perceptions, can revise beliefs in the direction of inferential and evidential consistency and generate new beliefs to explain and accommodate new information. The mental states involved in these sorts of cognitive activities require metarepresentation, the representation of the relation between a representation (essentially a propositional attitude) and its represented object. The flexibility and sophistication of human behaviour are based on such metarepresentational capacities. The novelty of EP is in its claim that some or all of these central metarepresentational processes are modular. Like peripheral modules they have specific domains and exploit specialised algorithms or data stores and may have localised neural substrates.

A crucial piece of evidence for the claim that central processes have a modular architecture is the Theory of Mind (TOM) module. The existence of a TOM module was first postulated by developmental psychologists to explain the difference between normal and autistic children on “false belief” tests (Baron-Cohen et al. 1985).¹ For EP these results are evidence that there is a module, whose domain is the metarepresentation of propositional attitudes, whose failure to come online in autistic children explains their problems in the domain of *social* cognition. EP has adopted the arguments of modular TOM theorists that the ability to understand other minds (“mindreading” as it is known) is not something we learn as the result of domain general cognitive processes. Rather our ability to do so is the result of maturation of a genetically specified module.¹ “Such inferences appear to be generated by a domain specific cognitive system that is sometimes called a ‘theory of mind’ module” (Cosmides and Tooby 1994, p.102).

I shall argue that mindreading is plausibly explained as the result of central, domain general abilities taking as inputs the modularised and peripheral processing which constitutes “early” social cognition. “Early” here refers both to ontogeny and cognition. Ontogenetically, peripheral abilities such as the recognition of intentional bodily or hand movement and emotional expression develop *before* higher order abilities such as propositional attitude attribution. Furthermore, since these type of modules provide *inputs* to higher order processing they constitute an *earlier* stage of social cognitive processing in the same way as recognition of motion is an earlier stage of visual processing than recognition of shape or colour, or the recognition of orientation is an earlier stage of face processing than recognition

¹ There are a number of variations of this task, but one simple version that has been used on autistic populations is the so-called “Sally-Ann” task (Baron-Cohen et al. 1985; Wimmer and Perner 1983): Children are shown two dolls, ‘Sally’ and ‘Ann’. Sally has a basket in which she places a marble. Then, she goes away leaving her basket behind. Ann takes Sally’s marble out of the basket and puts it in a box. Sally returns, and the children are asked: ‘where will Sally look for her marble?’ To pass, children must correctly predict that Sally will look in the basket where *she* believes the marble to be, as opposed to box where they know the marble is themselves.

of emotional expression. For example forming the propositional attitude that “ S is threatening” typically depends on earlier (in both senses) modularised responses to expressive body movement or vocalisation. It is absence of some or all these early abilities in autism which deprives autistic subjects of a crucial developmental resource and gives the misleading impression that the essential difference between autistic and normal subjects is at a higher level: a module concerned with social cognition. If this is correct EP has lost substantial support for one of its central claims, that the mind is modular at the centre as well as the periphery. In order to substantiate this claim we need to look more closely at the nature of modularity, the developmental findings, and the autistic deficits.

Modules, domain specificity and informational encapsulation

Within EP modules have been conceptualised in three ways: (Samuels 1998)

1. The hardware conception. The human brain is a set of cognitive devices with *distinct neural realisations*.

The brain must be composed of a large collection of circuits with different circuits specialised for solving different problems. One can think of each specialised circuit as a minicomputer that is dedicated to solving one problem. Such dedicated minicomputers are sometimes called modules. (Cosmides and Tooby (1997), p.81)

2. The algorithmic conception. Modules are individuated, not by their physical but their computational architecture.

This rich array of cognitive specialisation can be likened to a computer program with millions of lines of code and hundreds or thousands of *functionally specialised subroutines*. (Cosmides and Tooby 1992, p.39)

3. The epistemic conception. Modularity is a domain specific body of *innate knowledge*.²

Intuitive mechanics: knowledge of motions forces and deformations that objects undergo- Intuitive biology: understanding how plants and animals work . . . Intuitive psychology: predicting other people’s behaviour from their beliefs and desires. (Pinker 1994, p.420)

These conceptions of modularity are non-equivalent ways of making more precise

² Robert Samuels has pointed out that the epistemic conception is quite consistent with the idea that the algorithms and hardware involved in cognising the domain are non-modular. Perhaps innate knowledge is recruited by domain general cognitive processes, which would explain why certain abilities are developmentally precocious. For example if knowledge of grammar is innate this, rather than the presence of a functionally dedicated device or algorithm, would explain why knowledge of language comes so effortlessly and automatically. However we would still expect to see some developmental dissociations of language ability. On this view the mind is like a library some of whose shelves are prestocked. (Samuels 1998). But the retrieval and use of the information stored on the shelves follows essentially the same domain general cognitive procedures. (Gerrans, forthcoming). Where the shelves are not stocked we see a developmental dissociation.

the implications of the basic notion of modularity which is that cognition is *domain specific*, not domain general. Each domain is cognised by a dedicated module understood in either the hardware, algorithmic or epistemic conceptions outlined above. The first stage in argument for modularity is thus, typically, a quite general one in favour of domain specificity, which is then refined into one or other of the conceptions above. Which direction that refinement takes is dictated by the empirical evidence about neural and computational architecture involved in cognising the particular domain, or an argument that cognising the domain necessarily depends on an innately specified body of knowledge.

Samuels is right that EP has tended to ignore the distinctions between conceptions of modularity, either assuming their equivalence or downplaying their importance. This quotation from the Ur manifesto is an instance:

There are specialised systems for grammar induction [knowledge] face recognition [hardware/algorithm] for dead reckoning [algorithm] for construing objects [knowledge] and for recognising emotions from the face [algorithm/hardware]. There are mechanisms [hardware] to detect animacy [knowledge], eye direction [algorithm] and cheating [knowledge]. There is a 'theory of mind' module [knowledge] . . . a variety of social inference modules [knowledge] and multitude of other elegant machines [hardware] (Tooby and Cosmides (1995), p. xiv)

The manifesto conjoins the three conceptions in the idea that each module is a separate body of innate knowledge, exploiting distinct algorithms, implemented in distinct hardwares.

Our cognitive architecture resembles a confederation of thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems (Tooby and Cosmides (1995), p. xiii)

In the case of TOM, EP theorists vary. Baron-Cohen (1995) has defended the strongest view, that the TOM unites all three conceptions. Leslie (1987), Leslie and Thaiss (1992), Leslie (1994) has emphasised the algorithmic conception and, in some moods, Pinker (1994) and Sperber (1994) have placed emphasis on the idea the TOM is an innate body of knowledge. On any of these views, however, the TOM module is innately specified and cognises a specific domain: namely the explanation of behaviour via the attribution of psychological states.

Furthermore, on any view of modularity, the operation of a module *cannot be affected by central, domain general processes*. This is why visual illusions persist, despite knowledge that they are not veridical. This property of a module is known as *informational encapsulation* (Fodor 1983) and can be briefly expressed as the idea that modular cognitive processes *are not susceptible to top down influence by the propositional knowledge of the agent*. This accords with the idea that modules provide *inputs* to central processes but they don't perform central processes. Furthermore, the non-modifiability of modular processing by central processes of belief fixation explains why modules operate automatically on their proprietary

inputs. A classic case is the Heider and Simmel montage in which subjects cannot help but interpret geometrical shapes, which they know are meaningless, in intentional terms. These shapes automatically engage a module which detects expressive body movement and which cannot be overridden by propositional knowledge. (Heider and Simmel 1944)

Informational encapsulation, understood as the insulation of modular from central processing poses a problem the EP theorist who argues that central processes are modular. She cannot mean that such central modules exploit domain general computational processes but apply them to specific domains (the “library” conception) because that destroys the core idea that modules are informationally encapsulated devices. Rather she must mean that central processes, hitherto conceived of as a domain general set of computational procedures, actually comprise a set of more domain specific procedures. There is a language module, a TOM module, a cheater detection module and so on, each of which require higher order cognition of their domain but whose cognitive processing is autonomous. Thus the TOM module must be a metarepresentational cognitive subsystem which has *specialised procedures for computing all and only the information required to attribute psychological states*. This is indeed a sharp contrast to my claim that the appearance of modularity in TOM cases is produced by domain general cognition operating on inputs provided by a variety of modularised mechanisms concerned with different aspects of early social cognition. One way to determine the issue would be to see whether TOM performance is affected by domain general processing. If it can be strongly affected by the subject’s propositional knowledge *prima facie* we have violation of informational encapsulation and evidence against the modular hypothesis.

Theory of mind as a specific domain

Arguments in favour of domain specificity for the TOM come in two forms: general arguments for the evolution of innate modules and empirical evidence from experimental and developmental psychology which confirm the presence of a TOM module in our cognitive phenotype. According to Lela Cosmides and John Tooby, we might expect innately specified domain specificity in cognition for these reasons.

The human mind is a cognitive phenotype and, as with physical phenotypes, there are constraints on what types of phenotypes can evolve. Think for a moment of phenotypical traits, both physical and mental, as *specialised* responses to problems set by the environment and, therefore, independent phenotypical components. In fact it is mandatory to do so because if the phenotype were not composite, evolution would be unable to debug traits, one by one, over the generations. A mutation which produced a smarter, tool-using hominid who was unfortunately dumber at detecting the hostile intentions of his conspecifics would not result in the presence of smarter tool users in the next generation. In this respect the hominid mind is no different to the hominid pelvis or vocal cords. Because phenotypical traits are specialised solutions to adaptive problems traits have to be able to evolve in isolation. (Lewontin 1978). Thus we might expect higher-order cognition to exhibit some

partitioning which reflects the adaptation of cognition to specific problems posed by ancestral environments.

The second argument for modularity is that specialised cognitive devices avoid the frame problem and its consequences for real world learning (Fodor 1987; Sperber and Wilson 1996). Organisms with specialised cognitive devices do not have to search the total cognitive space of possible solutions when confronted with a problem. Nor do they have to try out different responses in the real world to check whether a behavioural solution is the right one. Thus, once again, where organisms are confronted with the same cognitive problem over generations we might expect cognitive traits to evolve as specialised adaptive responses to those problems.

So, if our ancestors were confronted with a stable set of problems whose solution required specialised higher order cognitive processes, we might expect domain specific forms of cognition to evolve. Another way to put this is to say that the mind *might* be modular at the centre as well as the periphery *if* the relevant evolvability constraints were present. (Pinker 1997; Plotkin 1997). Namely stable and structured environments which present discrete problems requiring higher order cognition and evolvable cognitive architectures.

In the case of TOM the hypothesised central module is a solution to the problem of inferring the intentions of other conspecifics. Clearly a primate who could not work out whether another had aggressive or deceptive intent would be at a disadvantage. Furthermore, the primate who was able to deceive others in virtue of their inability to detect his disguised intentions would be at a considerable advantage in situations of competition for mates and resources. Thus we might expect an arms race in the development of the ability to detect intentions (Whiten and Byrne 1997). Whether, however, the result of such an arms race in primate evolution is a domain specific TOM module is open to question. There is good reason to think that many of the abilities cited as evidence for primate TOM can be redescribed as instances of associative learning (a paradigm domain general cognitive capacity) of complex behavioural routines. (Heyes 1995, 1998, (in press); Povinelli 1996)

Here evolutionary considerations actually cut both ways. No doubt it would be fitness enhancing to be able to detect another's intentions, but should we really believe that all the information relevant to that task might be modularised? This point relates to the earlier discussion of informational encapsulation in the following way. In order to evolve an informationally-encapsulated mindreading system the relevant stimuli would have to be reliable. This is why honest or involuntary signalling, via facial expression and bodily posture seem to be a part of the primate emotional repertoire. (Frank 1988). However if responses to all social signals become informationally encapsulated, deception becomes too easy. A conspecific could be tricked again and again because unable to override her trusting responses to signals faked by her smarter companions. Thus a *modularised* capacity for detecting intentions would make one highly vulnerable to deception. All the deceiver would need to do would be to use a deceptive strategy whose recognition required information outside the cognitive parameters of the module. (Sterelny 1995). We might then expect an arms race in the direction of domain generality or demodularisation!

In the case of human mindreading the argument for demodularisation seems even more compelling. What information do you need to infer that I am lying? Could there be a *module* which represents all and only the information required to realise that someone does not normally get lipstick on the collar while working late? Is it the same module which helps you realise that the friendly colleague, talking to the boss about your work is actually trying to steal your job? In both these cases once one understands the beliefs and desires of one's informants and puts actions into a global context everything becomes clear. The most likely cognitive hypothesis here is that a global cognitive capacity for abduction and inductive confirmation is involved. (Bloom and German 2000). Of course you might be alerted to someone's lying intent by their posture, vocalisation and refusal to meet your gaze. But detecting these as significant uses *early* mechanisms of social cognition. The hypothesis that they are lying is an *explanation* of the feeling you get when these early mechanisms of social response are activated.

There is no doubt that chimpanzees and higher primates have modularised capacities for some *early* aspects of social cognition (as opposed to routines acquired by associative learning) but they are far more circumscribed than a modularised capacity for detecting intentions. For example response to faces, emotional expression and posture (Adolphs et al. 1994, 1998; Heberlein et al. 1998) and the recognition of goal directed action form part of the primate repertoire and are neurally and cognitively homologous in humans. For example Brodmann's area 6 in humans, implicated in the detection of intentional hand movements, is apparently an homologue of F4 and F5 in the macaque premotor cortex. (Gallese et al. 1998; Decety et al. 1994). It does seem clear, however, that primates and humans differ vastly in respect of their TOM abilities. For example, no primate has ever unambiguously passed the false belief test.

Modularity and innatism

Evolutionary considerations not being decisive, the strongest argument for a modular theory of TOM in humans rests on the developmental dissociation of TOM abilities in autism. The developmental aspect is crucial to the EP claim that the development of the TOM module is *innately constrained* in the same way as the development of (say) language ability or face recognition. This is important, because without a developmental explanation, the presence in the phenotype of specialised cognitive abilities and specialised neural structures is not sufficient evidence for a theory of innate mental modularity.

To see why, recall that modularity hypotheses were first developed for the automatization and routinisation of cognitive processing observed in cases of expertise. Expert musicians, sportsmen, chess players and readers may start out by consciously applying central domain-general forms of cognition and lower level modular processing, but, as their expertise develops, they build subroutines which automate the operations involved. So much so that it looks to the observer and feels to the subject as if the processing has become as effortless and automatic as

perceptual processing. In effect the expert has built a specialised subsystem by automating some central/ modular interactions. Furthermore, once modularised, these subsystems cannot be overridden by central processing. They are informationally encapsulated. (This is why it is so hard for sportsmen acting in realtime, not the practice field, to overcome old habits. The module they have painstakingly automated over years recognises its proprietary stimulus and goes into action. Typically, their more evolved and less stimulus bound opponent anticipates the result.)

So, even though there was selective pressure to become good detectors of others' intentions in the ancestral environment and the normal human develops mindreading abilities, *possibly* subserved by dedicated neural circuitry, these facts alone are not yet conclusive evidence that we have an innately specified mindreading module. The same abilities could be the result of the routinisation of the central-modular interface as part of normal development within an intensely social context.

In support of innatist hypotheses EP typically relies on two considerations. The first is the universality of the development of mindreading in the normal population. Here the most crucial piece of evidence is the age at which children pass the false belief test and its companions. These tests identify the ability to make behavioural predictions which, it is argued, necessitate the metarepresentation of intentional states. There are robust cross-cultural findings indicating that the normal window is between the age of three and half and four years. "Between the ages of 3 and 5 this domain specific inferential system develops in a characteristic pattern that has been replicated cross culturally in America, Europe, China Japan and Cameroon. Hence, just as the development of language has a typical trajectory, the coincidence of mindreading developmental trajectories is taken as evidence for the theory that we all share the same genetically-specified TOM competence. "This research suggests that a panhuman theory of mind module structures the folk psychology that people develop" (Cosmides and Tooby 1994, p102)

One crucial point here is a disanalogy between classic modular theories of language acquisition, on which the EP version of TOM is modelled, and mindreading. So called Poverty of Stimulus (POS) arguments point out that:

1. there are wide variations in the amount of linguistic exposure (Primary Linguistic Data. PLD) children receive.
2. even within any individual's PLD the amount of grammatical information received drastically underdetermines any hypothesis as to which of many possible grammars she is confronted with.

Yet, the argument goes, all children acquire their natural language automatically via the same developmental trajectory and converge on the same set of linguistic universals. They cope effortlessly with grammatical novelty, parsing and generating grammatically adequate sentences they have never confronted in the PLD. Hence the conclusion that some basic grammatical knowledge must be innate.

However the same considerations do not apply to the primary psychological data (PPD) confronting children. Modular EP theorists are captured by the problem posed by the fact that propositional attitudes are not directly perceptible, a fact dramatised in a paragraph of Alison Gopnik's which describes how a human would

look to someone who could not automatically and effortlessly interpret its behaviour as evidence of mental states:

Around me bags of skin are draped over chairs, and stuffed into pieces of cloth; They shift and protrude in unexpected ways- two dark spots near the top of them swivel restlessly back and forth. A hole beneath the spots fills with food and from it comes a stream of noises . . . (Gopnik 1993).

Gopnik's point is of course that we see people lounging, talking and eating, precisely because we know these bags of skin have minds. Similarly, we don't hear meaningless noises when people talk, we hear sentences. In both cases we make inferences to the unobserved mental causes of the primary data. We hear speech as language and see behaviour as action because we apply the requisite grammatical and intentional concepts automatically and effortlessly. The EP explanation is that the proprietary data (language or intentional behaviour) activates an innate module.

In the case of TOM, however, EP is too quick to assimilate POS arguments for innate modularity, valid in the linguistic case, with an argument that inference from observed to unobserved in *any* domain requires innate modularity. "To allow a human to represent at least some of the mental states that generate others' behaviour, special inference systems must be available to bridge the gap from the observable to the unobservable" (Cosmides and Tooby (1994), 102). Of course propositional attitudes are not observable, *but evidence licensing an inference to their presence is surely ubiquitous and universal*. The child's social environment is a relentless set of clues which seem to lead inexorably to the hypothesis that logically opaque mental representations cause behaviour. While it is certainly true that humans cannot form, in a single lifetime, a sufficient body of data concerning the results of incest to validate the hypothesis that it is maladaptive, infants pre-equipped with early mechanisms of social cognition and affective response do get a lot of evidence that people act on the basis of mental states. What other plausible hypothesis explains the data they receive? It is only psychologists and philosophers who have been tempted by alternatives such as behaviourism or eliminativism.

In fact the problem for the infant is not how to get from the situation described by Gopnik to being a competent mindreader, using a single module (which might indeed require a conceptually sophisticated theory of mind module). Rather the child builds on the repertoire of innately specified *early* mechanisms of social recognition and response to make the inference from observed behaviour to unobserved minds. Once one recognises this, the inference seems quite highly constrained: so much so that even domain general minds would converge on the hypothesis that people have intentions. Particularly when, for language using infants, their social environment is saturated with references to precisely such unobservable entities. The autistic child however is in something like the predicament described by Gopnik, because she lacks the early cognitive mechanisms of social cognition which transform her environment from an intentionally inert world to one alive with the possibility for emotionally rewarding social interaction.

Furthermore, as Cosmides and Tooby themselves point out, treated as a data set for the development of mindreading, the infant's environment is totally predictable. One does not have to be a hard-core developmental systems theorist to note that the

combination of multiple early mechanisms of social cognition, together with the intergenerational stabilities in the social and psychological environments would tend to canalise development in the direction of mindreading (Lewontin 1978; Griffiths and Gray 1994). For Cosmides and Tooby domain specificity requires that the mind have an innately specified device which “locks onto” proprietary data. In other words that the mechanism by which mindreading is transmitted must be genetically wired into to the mind of the agent. They seem to have overlooked the possibility that stable regularities in the environment may be part of the mechanism of inheritance.

Explaining the appearance of modularity in autism

The second argument for an innate TOM module is the performance of autistic subjects on false belief tests. Autism, understood as a failure to cognise others' mental states is taken as evidence of the dissociation of the cognitive ability involved, and hence symptomatic of a dedicated TOM module which subserves it. However, I shall suggest that autism is not a result of damage to a mindreading module but the result of neurological damage to a number of early cognitive mechanisms whose proper function is essential to the development of mindreading.

The absence of these inputs to social cognition may arise in two ways. Firstly as the result of a neurodevelopmental disorder which affects subsystems which are not specifically social but which are necessary to development of normal social cognition. Sensory, motor and affective responses are candidates and autistic subjects have abnormal profiles in all these areas. Secondly, as the result of a disorder which affects crucial early social processing such as the recognition of expressive body movement, goal directed action or emotional expression. And it is well documented that autistic subjects have deficits in these areas. (Baron-Cohen et al. 2000)

Autistic subjects share a distinctive triad of impairments in social, communicative and imaginative capacities (the latter demonstrated by the absence of pretend play in childhood and restricted interests and activities that persist throughout life) (Rutter and Schopler 1987; Wing and Gould 1978, 1979). Although seventy-five percent of diagnosed autistics are mentally handicapped in a general way (as reflected in low IQ scores), the remaining twenty-five percent have normal to high IQ's and often perform well, and sometimes better than average, on reasoning tasks that don't require any understanding of the mental life of agents. By contrast, on so-called 'theory of mind' tests, these “high-functioning” autistic children are significantly impaired when compared with normal children and even those with Downs Syndrome who are matched with them by mental age (for a review of research, see Baron-Cohen et al. (2000)). For instance, on first-order false-belief tasks, which require subjects to predict another's behaviour on the basis of attributing to them a false belief, children will normally pass by a mental age of four years (Wimmer and Perner 1983). Autistic subjects, if they pass at all, only do so when they are considerably older: on average, at a verbal mental age of nine (Happe 1994b).

A prime example of this dissociation between social and non-social reasoning skills involves the Zaitchik ‘false-photograph’ task, which is modelled on the standard false-belief task except in so far as it tests children’s ability to reason about physical (photographic) representation instead of mental representation (Zaitchek 1990). In one version of this task, children are shown how a Polaroid camera works. Then a picture is taken of a play scene in which a toy cat is sitting on a chair. The photograph is taken from the camera and put face down on a table. As the photograph develops, the experimenter changes the play scene by moving the cat from the chair to a nearby bed. Then, before turning the picture face up, the experimenter asks the children: “in the photo, where is the cat sitting?” High-functioning autistics who fail the false-belief task have no trouble answering this question correctly. They understand that the photograph will show the cat sitting on the chair and not on the bed (Leslie and Thaiss 1992). This pattern of failing false-belief while passing false-photograph tasks does not occur in normal four-year olds.³ Failure to pass the false belief test is only one quite late instance of the general mindblindness of autistic subjects from early infancy. For example they are emotionally unresponsive and withdrawn, do not track gaze, engage in normal pretend play or engage in social referencing.³

The TOM hypothesis unites all these all these failures together with the characteristic language and imaginative defects by treating them as failures of a single modularised capacity: the TOM module. Results like these are cited as evidence for the hypothesis that autism represents a developmental failure of “innate, isolable component of the mind which embodies a *theory* of the nature and the operations of mind” (Carruthers and Smith 1996, p. 258): a TOM module. Mindblindness explains the nature of the linguistic deficit which is restricted to pragmatic and non literal aspects of communication. It can also account for the lack of imaginative play and spontaneity in terms of inability to conceptualise action mentalistically. Perseverative and stereotyped autistic actions are thus remain primarily motor, not intentional, phenomena.

So, *prima facie*, the TOM is a very attractive unifying program, but we should note the autistic abnormalities that seem to have little to do with ‘theory of mind’ capacities. These include sensorimotor problems: e.g., extreme and unusual physical sensitivities and insensitivities; slowed orienting of attention; oddities of posture and gait; tics, twitches and unusual mannerisms; stereotypies such as rocking, hand-flapping, spinning, thumb-twiddling, and echolalia. They also include abnormalities in perceptual processing, leading to a characteristic autistic profile of assets and deficits on various perceptual tasks: e.g., insusceptibility to certain perceptual

³ Although some studies indicate normal children may show dissociation in the opposite direction (passing false-belief and failing false-photograph) (Leslie and Thaiss 1992; Zaitchek 1990), more recent studies suggest this may be an artefact of experimental design. Normal four-year olds do equally well on both tasks once incidental conversational and linguistic differences between them have been eliminated. autistic subjects, on the other hand, continue to show the dramatic physical-mental dissociation seen in earlier studies (Peterson and Siegal 1998; ??? 1990). The expository material on pages 10 and 11 concerning autistic deficits abridges and paraphrases (Mc Geer 2001) and Gerrans and McGreer (forthcoming).

illusions, superior performance on finding embedded figures within a larger design, superior visual memory and capacity for rendering scenes in precise detail, perfect pitch, difficulties with ‘gestalt’ perception – seeing whole figures or scenes as opposed to their parts, absence of perceptual ‘switching’ with ambiguous figures such as the duck-rabbit, and so on.

When we consider this range of symptoms a non cognitive explanation which treats autism as a syndrome cause by a low level brain disorder looks attractive. The possibility is illustrated by the metabolic disorder known as phenylketonuria (PKU). Although PKU has many symptoms that are similar to autism, it is produced, not by the failure of any cognitive module, but by the absence of an enzyme which synthesises an amino acid. (Page 2000) In this case the cause of the disorder is essentially a non-cognitive neuropathology which produces a wide range of symptoms, many identical with those of autism.

A TOM theorist might concede that autism is, in effect, a collection of relatively independent disorders resulting from multiple failures across a variety of distinct neurological systems but maintain the hypothesis that one such subsystem is the TOM module and it is this which explains the *specifically social deficits of autism*. It then becomes crucial to establish whether the mindreading failures of autistics require the TOM hypothesis or whether they can be explained as a consequence of the failure of early social processing, sensory, motor and affective processing.

Philosophers and cognitive psychologists have paid insufficient attention to the possible link between autistic sensory disturbances and autistic “mindblindness”, despite the suggestive label. If autism is partly a consequence of sensory disturbances, then it would seem that *other clinical populations with sensory problems ought to show similar kinds of higher-order TOM abnormalities*. And indeed this is the case. Deaf children of hearing parents as well as congenitally blind children suffer autistic-like deficits in social, communicative, and imaginative abilities, as well as selective incapacity to pass reasoning tasks with a mentalistic component (Brown et al. 1997; Hobson 1993b; Peterson et al. 2000; Peterson and Siegal 1998, 1999). Indeed, the parallels among these populations call for a unifying explanation. Naturally, however, that unifying explanation cannot be in terms of a modularised cognitive deficit shared by autistic, blind and deaf children, since these blind and deaf children are in all respects, other than their deafness, normal. Also blind and deaf children eventually develop normal mindreading once their sensory deprivation is compensated for.

The similarity between all three populations is absence of the type of early input required to sensitise a subject to the presence of other minds. Blind children cannot perceive goal directed action or expressive body movement for example. And they cannot monitor another’s gaze or recognize emotional expression. Autistic children are not the same as blind children but their disorder deprives them of some early inputs to social cognition such as responding to gaze and expressive body movement. In fact in their failure to react to expressive body movement autistic patients resemble frontal patients, and this has encouraged the idea that the orbito frontal cortex is the seat of the theory of mind module (Baron-Cohen and Ring 1994; Brown et al. 1997; Stone et al. 1998). Autistic like symptoms can also be produced

by amygdala damage. In a recent study of a patient with bilateral amygdala damage, Heberlein et. al. showed a patient the Heider and Simmel video, a montage of geometric shapes which normal subjects cannot help but describe in intentional vocabulary, attributing psychological states (fear, worry) roles (bullying) and narratives (chasing). The standard explanation is that the montage activates a modular capacity for responding to expressive body movement. For example the end of the video shows a large triangle “breaking” or “destroying” the rectangle (as typically explained by normal, and control brain-damaged subjects, who also use terms like “bullying” to describe the interaction of the geometric shapes). SM simply said “And then the two parts of the rectangle made like an upside down V, and that was it” (Heberlein et al. 1998). Autistic subjects exhibit the same type of deficit, not only in the geometric cases but with sequences of human behaviour.

Another crucial developmental resource absent in autism is a normal pattern of affective response to other humans. This is not to endorse a reductive affective theory of autism but to point out that crucial keys to early development such as imitation of smile and gesture depend on perceiving and reproducing the bodily expressed emotional states (Hobson 1991, 1993b; Stern 1985; Trevarthen 1979; Trevarthen and Hubley 1978). As McGeer has emphasised, affectively modulated interaction provides information about the world and human experience, and serves as a critical developmental resource. (Mc Geer 2001). This fits with a theme emphasized by Hobson and other theorists that an important component of normal social cognitive development is the affective quality of intersubjective experience (Hobson 1991, 1993b; Stern 1985; Trevarthen 1979; Trevarthen and Hubley 1978). If the infant’s sensory, affective perceptual and motor systems are abnormal her motivation and ability to engage in the type of behaviour necessary to acquire mind reading and will be disrupted from the very beginning (Gerrans 1998)

The case of the deaf children observed by Petersen reminds us that sensory, affective, and early social cognitive responses to other people are not the only route to other minds. As well as normal sensory experience and early social cognition, deaf children are also deprived of another crucial developmental resource. Namely, ready access, reinforced by sensory evidence and social interaction, to *intentional concepts*. Petersen conjectured that the primary deficiency experienced by her non-signing deaf children was the use and discussion of intentional vocabulary. Reference to higher order intentional states is a conceptually sophisticated task for which forms of communication simpler than fully-fledged signing are inadequate. Non-signing children are not exposed to discussion of others’ motives employing second order intentional concepts and hence they are unable to use them to interpret situations, such as the false belief task, which require them. When those children acquire the concepts, via sign language, they pass the false belief test.

This tends to suggest that the deaf children are learning to apply psychological concepts in social contexts using general metarepresentational abilities. When this intellectual superstructure is added to the range of scaffolding already present in the variety of sensory and recognitional mechanisms employed in social interaction, the deaf children, who are otherwise normally embedded in their social world, can employ them with the normal facility. Thus language, as well as normal sensory

awareness and social engagement is crucial to passing the false belief test because it is the primary mode of gathering the requisite concepts. The role of language and general cognition is further highlighted by the fact that normal three year olds can be coached to pass the false belief test if they receive intensive exposure to metapsychological concepts.

Conclusion

There is no innate theory of mind module. Of course many of the *cognitive prerequisites* for mindreading, such as mechanisms for the detection of emotional expression and goal directed behaviour are genetically specified and plausible candidates for modularity hypotheses. But that is precisely because these cognitive traits are both domain specific and informationally encapsulated. They develop independently of central processes and cannot be affected by them. Mindreading is learning to synthesise these inputs to discover what someone else is thinking in a social context (Waterhouse et al. 1996) And the hypothesis, based on the case of autism, that this learning process is insulated from central forms of cognition seems unwarranted. One can explain autism on the assumption that the domain specificity of the deficit is the result of early deficits influencing development. Furthermore, the more general arguments made by EP theorists that we need to postulate a module to explain why normal children converge on the hypothesis that unobservable mental states explain behaviour are not decisive. The evidence is staring the infant in the face.

References

- Adolphs R., Tranel D., Damasio H. and Damasio A. 1994. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* 372: 372–669.
- Adolphs R., Tranel D. and Damasio A. 1998. The human amygdala in social judgement. *Nature* 393: 470–474.
- Baron-Cohen S., Leslie A.M. and Frith U. 1985. Does the Autistic Child have a Theory of Mind? *Cognition* 21: 37–46.
- Baron-Cohen S., Tager-Flusberg H. and Cohen D.J. (eds) 2000. *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*. Oxford University Press, Oxford.
- Baron-Cohen S. and Ring H. 1994. A model of the mindreading system. *Neuropsychological and neurobiological perspectives*. In: Mitchell P. and Lewis C. (eds), *Origins of an understanding of mind*. Hillsdale, Erlbaum, NJ.
- Baron-Cohen S., Ring H., Wheelwright S., Bullmore E., Brammer M., Simons A. et al. 2000. Social intelligence in the normal and autistic brain: an MRI study. *European Journal of Neuroscience* 11: 1891–1898.
- Baron-Cohen S. 1995. *Mindblindness: An Essay on Autism and the Theory of Mind*. MIT Press, Cambridge, MA, pp xi-xviii.
- Bloom P. and German T. 2000. Two reasons to abandon the false belief task as a theory of mind. *Cognition* 77: B25–B31.
- Brown R., Hobson R.P. and Lee A. 1997. Are There “Autistic-like” Features in Congenitally Blind Children? *Journal of Child Psychology and Psychiatry* 38: 693–703.

- Carruthers P. and Smith P.K. (eds) 1996. *Theories of Theories of Mind*. Cambridge University press, New York.
- Coltheart M. 2000. Assumptions and Methods in Neuropsychology. In: Wixted J. (ed.), *Stevens' Handbook of Experimental Psychology. Methodology Vol. 4*. 3rd edn. .
- Cosmides L. and Tooby J. 1992. Cognitive Adaptations for Social Exchange. In: Dupre J. (ed.), *The latest on the Best*. MIT Press, Cambridge, MA.
- Cosmides L. and Tooby J. 1994. Origins of Domain Specificity. In: Hirschfeld L. and Gelman Susan (eds), *Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge.
- Cosmides L. and Tooby J. 1997. The Modular Nature of Human intelligence. In: Scheibel A. and Schopf J. (eds), *The Origin and Evolution of Intelligence*. Jones and Bartlett publishers, Sudbury, MA, pp. 71–101.
- Decety J., Perani D., Jeannerod M., Bettinardi V., Tadary B., Woods B. et al. 1994. Mapping Motor representations with PET. *Nature* 371: 600–602.
- Frank 1988. *Passions within reason: The strategic role of the emotions*. Norton, New York.
- Fodor J. 1983. *The Modularity of Mind*. MIT Press, Cambridge, MA.
- Fodor J. 1987. Frames Fridgeons Sleeping Dogs and the Music of the Spheres. In: Pylyshyn Z. and Norwood N.J. (eds), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Ablex.
- Gallese V., Fadiga L., Fogassi G. and Rizzolatti 1998. Action recognition in the premotor cortex. *Brain* 119: 593–609.
- Gerrans P. 1998. The Norms of Cognitive Development. *Mind and Language* 13: 56–75.
- Gerrans P. Rethinking Modularity. *Journal of Language and Communication* (forthcoming).
- Gopnik A. 1993. *Mindblindness*. University of California, Berkeley, CA. (unpublished).
- Griffiths P. and Gray R. 1994. Developmental Systems and Evolutionary Explanation. *Journal of Philosophy* 91: 277–304.
- Happe F.G. 1994b. An Advanced Test of Theory of Mind: Understanding of Story Characters' Thoughts and Feelings by Able Autistic, Mentally Handicapped and Normal Children and Adults. *Journal of Autism and Developmental Disorders* 24: 129–154.
- Heberlein A.S., Adolphs R., Tranel D., Kemmerer D., Anderson S. and Damasio A.R. 1998. Impaired attribution of social meanings to abstract dynamic visual patterns following damage to the amygdala. *Society of Neuroscience Abstracts* 24: 1176.
- Heider F. and Simmel M. 1944. An experimental study of apparent behaviour. *American Journal of Psychology* 57: 243–259.
- Heyes C.M. 1995. Knowing minds. Review of S Baron-Cohen 'Mindblindness', and D Byrne 'The Thinking Ape'. *Nature* 375: 290–290.
- Heyes C.M. 1998. Theory of mind in nonhuman primates. *Behavioral and Brain Sciences* 21: 101–148.
- Heyes C.M. Theory of mind and other domain-specific hypotheses. Author's Response to Continuing Commentary. *Behavioral and Brain Sciences* (in press).
- Hobson R.P. 1991. Through Feeling and Sight to Self and Symbol. In: Neisser U. (ed.), *Ecological and Interpersonal Knowledge of the Self*. Cambridge University Press, New York.
- Hobson R.P. 1992. Social Perception in High-Level Autism. In: Schopler E. and Mesibov G. (eds), *High-Functioning Individuals with Autism*. Plenum Press, New York.
- Hobson R.P. 1993a. *Autism and the Development of Mind*. Lawrence Erlbaum Associates Ltd., East Sussex, UK.
- Hobson R.P. 1993b. Understanding Persons: The Role of Affect. In: Baron-Cohen S., Tager-Flusberg H. and Cohen D.J. (eds), *Understanding Other Minds: Perspectives from Autism*. Oxford University Press, Oxford, pp. 204–227.
- Karmiloff-Smith A. 1998. Development itself is the key to understanding mental disorders. *Trends in Cognitive Sciences* 2: 389–399.
- Leslie A.M. 1987. Pretense and Representation. The Origins of a "Theory of Mind". *Psychological Review* 1987 94: 412–426.
- Leslie A. and Thaiss L. 1992. Domain specificity in conceptual development: evidence from autism. *Cognition* 43: 225–251.
- Leslie A.M. 1994. ToMM, ToBY and Agency: Core Architecture and Domain Specificity. In: Hirschfeld

- L. and Carey S. (eds), *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge.
- Lewontin R. 1978. Adaptation. *Scientific American* 239: 156–169.
- Mc Geer V. 2001. Psycho-Practice, Psycho-Theory, and the Contrastive Case of Autism: How Practices of Mind Become Second Nature. *Journal of Consciousness studies*.
- Oyama S. 1985. *The Ontogeny of Information*. Cambridge University Press, Cambridge.
- Page T. 2000. Metabolic approaches to the treatment of autistic spectrum disorders. *Journal of Autistic Developmental Disorders* 30: 463–469.
- Peterson C.C., Peterson J.C. and Webb J. 2000. Factors Influencing the Development of a Theory of Mind in Blind Children. *The British Psychological Society* 18.
- Peterson C.C. and Siegal M. 1998. Changing Focus on the Representational Mind: Concepts of False Photos, False Drawings and False Beliefs in Deaf, Autistic and Normal Children. *British Journal of Developmental Psychology* 16: 301–320.
- Peterson C.C. and Siegal M. 1999. Insights into Theory of Mind from Deafness and Autism. *Mind and Language* 15: 77–99.
- Pinker S. 1994. *The Language Instinct*. William Morrow & Co., New York.
- Pinker S. 1997. *How the Mind Works*. Harmondsworth, Penguin.
- Plotkin H. 1997. *Evolution in Mind*. London Allen Lane.
- Povinelli D. 1996. *What young chimpanzees know about seeing*. University of Chicago Press, Chicago, Ill.
- Rutter M. and Schopler E. 1987. Autism and Pervasive Developmental Disorders: Conceptual and Diagnostic Issues. *Journal of Autism and Developmental Disorders* 17: 159–186.
- Samuels R. 1998. Evolutionary Psychology and the Massive Modularity Hypothesis. *British Journal for the Philosophy of Science* 49: 575–602.
- Sperber D. 1994. *Explaining Culture*. Oxford Blackwell.
- Sperber D. and Wilson D. 1996. Fodor's Frame Problem and Relevance Theory (reply to Chiappe & Kukla). In: (ed.), *In Behavioral and Brain Sciences* 19., pp. 530–532.
- Stern D. 1985. *The Interpersonal World of the Infant*. Basic Books, New York.
- Sterelny K. 1995. The Adapted Mind. *Biology and Philosophy* 10: 365–380.
- Stone V.E., Baron-Cohen S. and Knight R.T. 1998. Frontal Lobe contributions to theory of mind. *Journal of Cognitive Neuroscience* 10: 640–656.
- Suddendorf T. and Whiten A. Mental evolution and development: evidence for secondary representation in children, great apes and other animals. *Psychological Bulletin* (in press).
- Suddendorf T. 1998. Simpler for evolution: Secondary representation in apes, children, and ancestors. *Behavioral and Brain Sciences* 21: 131–131.
- Tooby J. and Cosmides L. 1992. The Psychological Foundations of Culture. In: Barkow J., Cosmides L. and Tooby J. (eds), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, Oxford.
- Tooby J. and Cosmides L. 1995. Forward. In: (ed.), *Mindblindness: An Essay on Autism and the Theory of Mind*. MIT Press, Cambridge, MA pp xi–xviii.
- Trevarthen C. 1979. Communication and Cooperation in Early Infancy: A Description of Primary Intersubjectivity. In: Bullowa M. (ed.), *Before Speech: The Beginning of Interpersonal Communication*. Cambridge University Press, New York.
- Trevarthen C. and Hubley P. 1978. Secondary Intersubjectivity: Confidence, Confiding and Acts of Meaning in the First Year. In: Lock A. (ed.), *Action, Gesture and Symbol: The Emergence of Language*. Academic Press, London.
- Zaitchik D. 1990. When Representations Conflict with Reality: The Preschooler's Problem with False Belief and 'False' Photographs. *Cognition* 35: 41–68.
- Whiten A. and Byrne (eds) 1997. *Machiavellian Intelligence. Extensions and Evaluations*. Cambridge University Press, Cambridge.
- Wimmer H. and Perner J. 1983. Beliefs about Beliefs: Representation and the Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* 13: 103–128.
- Wing L. and Gould J. 1978. Systematic Recording of Behaviours and Skills of Retarded and Psychotic Children. *Journal of Autism and Childhood Schizophrenia* 8: 79–97.

- Wing L. and Gould J. 1979. Severe impairments of social interactions and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders* 9: 11–29.
- Winner E., Brownell H., Happe F., Blum A. and Pincus D. 1988. Distinguishing lies from jokes- theory of mind deficits and discourse interpretation in right hemisphere brain damaged patients. *Brain & Language* 62: 89–106.
- Zaitchek D. 1990. When representations conflict with reality: the preschooler's problem with false belief and "false" photographs. *Cognition* 35: 45–57.
- Waterhouse L., Fein D. and Modahl C. 1996. Neurofunctional mechanisms in autism. *Psychological Review* 103: 457–489.