Theories of theories of mind

edited by

Peter Carruthers

Professor of Philosophy and Director, Hang Seng Centre for Cognitive Studies, University of Sheffield

and

Peter K. Smith

Professor of Psychology, University of Sheffield

Published in association with the Hang Seng Centre for Cognitive Studies, University of Sheffield

CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

© Cambridge University Press, 1996

First published 1996

Printed in Great Britain at the University Press, Cambridge

A catalogue record for this book is available from the British Library

Library of Congress cataloguing in publication data

Theories of theories of mind / edited by Peter Carruthers
and Peter K. Smith
p. cm.
Chiefly papers presented at a series of workshops held in 1992 to
July 1994.
Includes bibliographical references and index.
ISBN 0 521 55110 2 (hardback) - ISBN 0 521 55916 2 (paperback)
1. Philosophy of mind–Congresses. 2. Philosophy of mind in
children–Congresses. 3. Cognition in children–Congresses.
I. Carruthers, Peter, 1952– . II. Smith, Peter K.
BD418.3T44 1996
128'.2–dc20 95-14610 CIP

ISBN 0 521 55110 2 hardback
ISBN 0 521 55916 2 paperback

10 The relationship between SAM and ToMM: two hypotheses

Simon Baron-Cohen and John Swettenham

1 Introduction

One of the most important achievements of modern developmental psychology has been to draw attention to the universal and astonishing capacity of young children to mind-read: it appears incontrovertible that by four years of age, children interpret behaviour in terms of agents' mental states (Wellman, 1990; Astington *et al.*, 1988). In John Morton's (1989) chilling phrase, they *mentalise*: they convert the behaviour they see others perform, or that they perform themselves, into actions driven by beliefs, desires, intentions, hopes, knowledge, imagination, pretence, deceit, and so on. Behaviour is instantly, even automatically, interpreted in terms of what the agent might be thinking, or planning, or wanting. What makes this developmental achievement of such interest is that it raises a learnability problem: how on earth can young children master such abstract concepts as belief (and false belief) with such ease, and at roughly the same time the world over? After all, mental states are unobservable, and have complex logical properties, as Leslie (1987), following Frege (1892) points out. If anything, we should have expected that mental state concepts should be bafflingly difficult to acquire, and yet even the most unremarkable child seems to understand them – without any explicit teaching. Reading minds seems to come naturally, whilst reading words seems to require a considerable amount of instruction.

Chomsky's (1965) solution to a similar learnability problem in relation to the acquisition of syntax was to postulate an innate mechanism or set of mechanisms dedicated to syntactic development; and as Pinker (1994) points out, children with specific language impairment, which can be of genetic origin (Bishop *et al.*, in press), may be the tragic but important evidence that such mechanisms not only exist but can be selectively damaged. In a similar vein, Leslie's (1991) solution to the question of how mind-reading is universally acquired was to postulate an innate mechanism, called ToMM (or the Theory of Mind Mechanism). This proposal gains considerable credibility from the evidence that children with autism have

selective impairments in the acquisition of mental state concepts, and in mind-reading (Baron-Cohen *et al.*, 1985; see Baron-Cohen *et al.*, 1993). They are in a real sense *mind-blind* (Baron-Cohen, 1990, 1995a). Given the innate basis of autism (Folstein and Rutter, 1988), one strong possibility is that genetic mechanisms normally enable neural structures or processes for mind-reading, and that in autism genetic abnormalities delay or prevent the development of such neurocognitive mechanisms. This genetic hypothesis is made more plausible by the evident adaptive value of mind-reading. Essentially, mind-reading allows flexible social interaction (based on shared plans), flexible communication (based on inferring a speaker's intentions, and sharing information that another individual might lack) and machiavellian deception (based on manipulating other's thoughts). As such, it is likely that mind-reading is a product of natural selection (Whiten, 1991).

In this chapter, we discuss ToMM in relation to an earlier developing mechanism: SAM, or the Shared Attention Mechanism. SAM is introduced in order to account for the developmental origins of ToMM. In particular, we examine the claim that SAM is a *necessary precursor* to ToMM. In doing so, we draw on two sorts of evidence: results from experimental investigations of autism, and results of a longitudinal study of infant behaviour predicting autism. Along the way, and following Gómez (1991), we unpack the precursor claim into two alternative hypotheses. But first we must introduce SAM.

2 SAM

SAM, or the Shared Attention Mechanism, is a special purpose neurocognitive mechanism, the function of which is to identify if you and another organism are both attending to the same object or event. SAM was first postulated by Baron-Cohen (1994),[1] as one component mechanism in the developing Mind-reading System in human beings. It is necessary to postulate a mechanism of this sort in order to explain how the infant transcends a purely solipsistic view of the world. SAM is quite a complicated but fundamental mechanism: it involves representing not only what another person sees (or wants), and not only what the self sees (or wants), but whether the self and the another person see (or want) *the very same thing*. We begin by briefly reviewing SAM's key features.

SAM develops in the normal human being at around 9–14 months of age. It builds 'triadic representations' which explicitly specify if you and another agent are attending to the same thing.[2] Triadic representations allow joint attention behaviours such as 'gaze monitoring' (when the child turns to look at the same object or event that another person is looking at

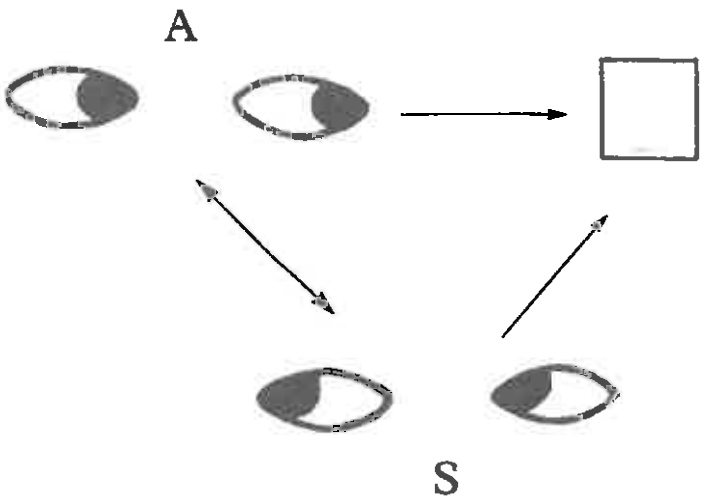

Figure 10.1 A triadic representation, expressed in pictorial format (reproduced from Baron-Cohen 1994, with permission)

– Scaife and Bruner, 1975; Butterworth, 1991); and 'protodeclarative pointing' (when the child uses a gesture (typically the outstretched index finger) to direct someone to attend to an object or event of interest, for its own sake – Bates *et al.*, 1979).[3] Triadic representations have the following form, expressed in a 'sentence'-like format: *[Agent/Self-Relation-(Agent/Self-Relation-'Proposition')]*. For example, *[I-see-(Mummy-sees-'the cup is on the table')]*.[4] An example of a triadic representation employed when the infant produces the protodeclarative pointing gesture is *[Mummy-sees-(I-see-'the cup is on the table')]*. Triadic representations, thus defined, are pretty complex, but then so is shared attention itself.

The examples above relate only to shared *visual* attention. However, SAM is amodal, meaning that it can build its triadic representations in any modality (visual, auditory, tactile). An example of a triadic representation built in the tactile modality would be *[I-touch-(Mummy-touching-'the cup on the table')]*. In practice, since building triadic representations in the visual modality is so much easier, and allows such a large number of potential objects or events to be shared in attention, SAM has a priority to be used in the visual modality.

In 'pictorial' form, triadic representations (in the visual modality) resemble figure 10.1. The arrows in figure 10.1 represent the relation term, and the two pairs of eyes represent the two agents (Self and another Agent). Note that the relation terms between Agents are in principle bi-directional

(hence the use of the double-headed arrow), whereas the relation term between an agent and an object is unidirectional. This specifies that only agents (but not inanimate objects) are capable of having a perceptual relation with something else.

The notion is that SAM constructs triadic representations out of simpler, 'dyadic' representations. Dyadic representations have the form *[Agent-Relation-'Proposition']*, for example *[Mummy-sees-'the cup is on the table']*. Dyadic representations are built by other, more primitive mechanisms. (These are not discussed here, but see Baron-Cohen, 1994, or 1995a.) These allow the infant to represent perceptual and volitional states, for example *[Mummy-wants-'the cup on the table']*. SAM thus not only builds triadic representations using perception terms, but also using desire/goal terms. As such, it equips the child with an 'attention-goal' psychology (Baron-Cohen, 1993). This also allows the child to read gaze direction in terms of volition (Baron-Cohen *et al.*, in press). Note that the mental state concepts that SAM processes are not *fully* intentional. For example, when the normal toddler of eighteen months represents that s/he and another person are *looking* at the same object, they do not at the same time represent their own and the other's *knowledge* states.[5] Nevertheless, this enables the child to build a simple, limited, but usable theory of mind.

SAM's development appears to be universal/independent of culture: children the world over show gaze-monitoring and protodeclarative pointing (Bruner, 1983). This implies that its development is partly, if not completely, driven by individual, biological factors within the child. Finally, SAM is held to be necessary (though not sufficient) for the development of the normal child's mature theory of mind. To expand this last point, we need to review the mechanism thought to be responsible for this.

3 ToMM

Normal children show a remarkable ability to understand a range of mental states, and to use such mental state concepts in making sense of and predicting action (Wellman, 1990). For example, at two years old they clearly understand pretending (Leslie, 1987; Harris, this volume) and desire (Wellman, 1990); at three years old they clearly understand that people have thoughts and know things (Pratt and Bryant, 1990; Wellman and Estes, 1986); and at four years old they clearly understand that people can have different (and even false) beliefs about the same state of affairs (Wimmer and Perner, 1983). As mentioned earlier, Leslie (1987, 1994a) suggests that a special purpose neurocognitive mechanism is responsible for the normal child's rapid, fluent acquisition of this mental state knowledge. This is

ToMM, or the Theory of Mind Mechanism. Here we briefly review ToMM's key features.

ToMM develops in the normal child around 18–24 months. Its earliest manifestation is in the production of pretend play, which Leslie (1987) argues involves the child representing its own or someone else's attitude or 'informational relation' (pretending) to a proposition.[6] To achieve this, ToMM employs 'M-Representations' which explicitly specify an agent's informational relation towards a proposition.[7] For example, suppose we hear John say 'I have a nice new hat.' We need to compute his meaning. Was he being truthful, or was he intending to deceive? Might he have only intended it as a joke? Was he being sarcastic? Or was he just pretending? Maybe he was meaning none of these things: maybe he was simply suffering from a mistaken belief. M-Representations allow us to represent what we, as a listener or observer, think John's attitude towards the utterance was.

According to Leslie, M-Representations have the following form, expressed in a 'sentence'-like format: *[Agent-Attitude-'Proposition']*. For example, *[John-pretends-'I have a nice new hat']*. (He might be putting a book on his head as he utters this.) The attitude slot in M-Representations can be filled by any intentional term (think, know, believe, intend, hope, warn, promise, wish, dream, wonder, imagine, etc.).[8] ToMM's M-Representations allow for a special logical property of *fully* intentional terms, namely, their non-substitutability (Frege, 1892). That is, substitution of identical terms in a proposition preceded by a fully intentional mental state term is not guaranteed to preserve the truth-value of the sentence as a whole. For example, if 'A man was killed in Oxford Street today' is true, and the man that was killed was actually John's father, then 'John's father was killed in Oxford Street today' must also be true. However, if 'John knows a man was killed in Oxford Street today' is true, it does not follow that 'John knows his father was killed in Oxford Street today' is also true. That will depend on whether John knows his father was in Oxford Street at the time.[9] (This contrast with SAM's ability to represent mental state terms that are only partly intentional is important.)

ToMM not only builds M-Representations, but also integrates the child's knowledge about the relationship between mental states and action into a mature and usable 'theory', to make sense of and predict another's action. ToMM's development also appears to be universal/independent of culture: children and adults, the world over, interpret behaviour in terms of mental states (Wellman, 1990; Fodor, 1983). This implies that its development is partly, if not completely, driven by individual, biological factors within the child.

4 The relation between SAM and ToMM: two hypotheses

In the section on SAM above, the final claim made was that SAM is necessary for ToMM's development. Elsewhere this relationship has been discussed as a causal relationship (Baron-Cohen, 1994), as SAM being a 'precursor' to ToMM (Baron-Cohen, 1989b), and as SAM 'triggering' ToMM to function (Baron-Cohen and Cross, 1992). In this part of the chapter, we wish to flesh out this idea a bit further.

Historically, one key impetus for the idea that SAM might have anything to do with ToMM was the experimental evidence that children with autism appear to be severely impaired in both domains – joint attention and theory of mind. Thus, regarding joint attention, children with autism do not show spontaneous gaze-monitoring (Leekam *et al.*, 1994) or spontaneous protodeclarative pointing (Baron-Cohen, 1989b) – indeed, the whole range of joint attention behaviours is impaired in autism (Sigman *et al.*, 1986).

Regarding theory of mind, children with autism do not produce much (if any) spontaneous pretend play (Baron-Cohen, 1987), and have inordinate difficulty understanding false belief (Baron-Cohen *et al.*, 1985; Perner *et al.*, 1989). Nor do they easily understand knowledge formation (Baron-Cohen and Goodhart, 1994) or the appearance-reality distinction (Baron-Cohen, 1989c), or show a normal understanding of intentions (Phillips *et al.*, forthcoming). They also produce few mental state terms in their spontaneous speech, relative to both normally developing children and those with a mental handicap (Tager-Flusberg, 1993). Finally, they show poor understanding of deception (Sodian and Frith, 1992; Baron-Cohen, 1992), and many even fail to make the mental-physical distinction (Baron-Cohen, 1989c). This literature is reviewed more fully in Baron-Cohen (1993, 1995a) and discussed critically in the volume edited by Baron-Cohen, Tager-Flusberg, and Cohen (1993).

4.1 The lock and key hypothesis

Since SAM produces triadic representations, it follows that if SAM triggers ToMM in some way, it must be because triadic representations have some special property to enable this. Furthermore, given that ToMM builds M-Representations, it follows that if SAM triggers ToMM, it must be because triadic representations 'activate' M-Representations in some way, since producing triadic representations is all that SAM does. We will refer to this as the 'lock and key hypothesis'.

The lock and key hypothesis suggests that triadic representations are the key, and M-Representations are the lock.[10] To see how this might work we need to examine the different parts of these two types of representation,

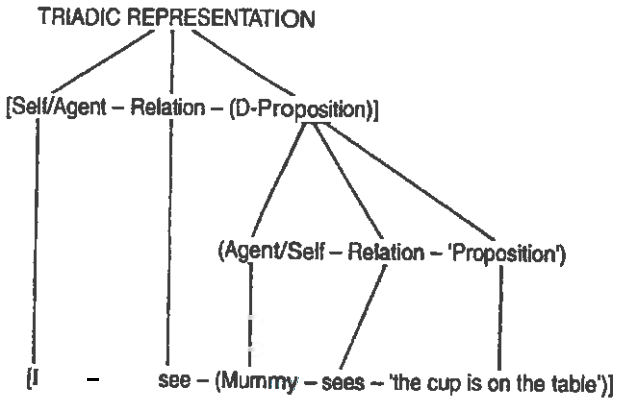

Figure 10.2 Tree structure of a triadic representation

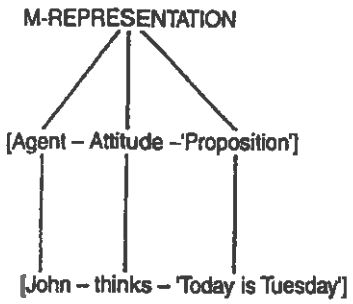

Figure 10.3 Tree structure of an M-representation

and identify in what way they may be 'naturally' compatible. The first thing to note is that triadic representations are like transitive-verb sentences: they have the unusual property of taking as their object an embedded proposition of an agent standing in a perceptual (or volitional) relation to a proposition. This is shown most clearly in a tree-structure analysis (fig 10.2). Here, we have called the special embedded proposition a D-Proposition, to mark that this is a proposition about an agent's *dyadic relation* to an object. It is not just any old proposition. The agent's dyadic relation may be perceptual or volitional. To continue with the tree-structure analysis, an M-Representation has the structure shown in figure 10.3, above.

If triadic representations are to be thought of as inputs to trigger M-

Representations, how might this happen? One possibility is that since both D-propositions (in triadic representations) and M-Representations are tripartite, they might have a special 'fit'. Another possibility is that cases where the triadic representation specifies that you and another agent are *not* attending to the same object/event may serve to distinguish the D-proposition in a triadic representation. For example, consider the triadic representation *[I-see-(Mummy-is not seeing-'the cup is on the table')]* (because I see she is looking in the opposite direction): such cases provide clear evidence that I and another agent can have different perceptual relations towards a state of affairs. D-propositions appear to have a compatible format to act as input for M-Representations (they have the virtually identical tripartite structure); and it may be that when they are within triadic representations they are clearly marked as belonging to one agent and not another, and can thus trigger M-Representations.

4.2 *The metamorphosis hypothesis*

An alternative to the lock and key hypothesis is that SAM and ToMM are not independent mechanisms; rather, SAM may be a developmentally earlier form of ToMM. We will refer to this as the 'metamorphosis hypothesis'. The metamorphosis hypothesis claims that, with development, SAM's triadic representations simply become more complex. For example, instead of only filling the Relation slot with *partially* intentional terms like *see*, *look*, *attend*, or *want*, the Relation slot can now be filled with *fully* intentional terms like *pretend*, *think*, *know*, *imagine*, and *believe*, etc., On this account, M-Representations are also triadic in structure, but use more complex mental state concepts.[11]

For this hypothesis to be plausible, we suggest that the definition of M-Representations be slightly revised, as follows: *(Self-Attitude-[Agent-Attitude-'Proposition'])*. An example of this would be *(I-think-[John-pretends-'the banana is a telephone'])*. This modified definition of an M-Representation allows the respective attitudes of Self and another Agent to be specified, in a parallel way to that seen in a triadic representation. Furthermore, this modified definition allows for the child to be aware that his or her own attitude might be different to that held by the other Agent.

Of course, saying that SAM's triadic representations have 'developed' into ToMM's M-Representations still begs the question as to how this has happened. Leaving this problem aside here, the key difference to bring out is that this alternative hypothesis suggests that there are not two independent mechanisms, but rather one mechanism changing in scope.

5 Testing between these two alternatives

In the previous section we have addressed the problem of analysing why triadic representations may be the right sort of input for activating M-Representations. To reiterate, the lock and key hypothesis is that SAM and ToMM are *independent* mechanisms which share a special developmental connection. The metamorphosis hypothesis is that SAM is a developmentally less mature form of ToMM. Both hypotheses predict that some children with autism may be impaired in SAM and therefore inevitably in ToMM, whilst late onset cases may have SAM intact but have an intrinsic impairment in ToMM. Yet other subgroups can be envisaged: those who are not impaired in these mechanisms in an all-or-none way, but who are significantly *delayed* in the functioning of one of them. Further research is needed to attempt to identify if such subgroups exist. As far as we can see, the only way of directly testing between these two hypotheses would be to examine cases of acquired damage (later in life) leading to double dissociations. Only the lock and key hypothesis would predict that double dissociations would be possible, and this would be strong evidence of the independence of these two mechanisms.

6 Evidence from early screening for autism

We wish now to switch to new evidence from our screening study of autism in infancy (Baron-Cohen, Cox, Baird, Swettenham, Nightingale, Morgan, Drew, and Charman, 1994). Whilst such evidence cannot test the two earlier hypotheses against each other, such evidence is relevant to the issue of whether SAM is necessary for the development of ToMM. This study is briefly reviewed next.

Between May 1992 and May 1993, 16,000 randomly selected children in the South East of England were screened at 18 months of age by their health visitors, using a specially designed checklist (called the *Checklist for Autism in Toddlers*, or the CHAT)[12]. Children with severe developmental delay were excluded from this population study. The key items in this checklist are protodeclarative pointing, gaze monitoring,[13] and pretend play. These are recorded by the parents and by the health visitor, separately. Protodeclarative pointing and gaze monitoring are of course joint attention behaviours, and thus should require SAM. Pretend play, according to Leslie, is an early manifestation of ToMM.

Of the 16,000 children screened, just twelve children failed all three items, on two administrations of the CHAT. Of these twelve cases, ten (or 83.3%) received a diagnosis of autism, using standardised diagnostic measures and established criteria. This is clear evidence that in autism there are

severe impairments in SAM and ToMM, even by 18 months of age. Furthermore, it suggests that these are important indicators in the early diagnosis of autism. Of relevance to the precursor issue, there were no cases, in all the 16,000 children, of children who failed both of the joint attention items but who unambiguously passed pretend play. We take this as clear evidence that if SAM is impaired, then ToMM will inevitably be too.

From the population, we picked out for detailed study a comparison group of 22 children who were not producing protodeclarative pointing at 18 months, but who did show gaze-monitoring, implying that they did not lack joint-attention completely. We can therefore think of these cases as being delayed in aspects of joint attention, but not being severely impaired in SAM. From other work, we can predict that such children will be at risk not for autism but for forms of developmental delay (Tomasello, 1988). In our study, 15 of these 22 cases received a diagnosis of developmental delay, and none of these 22 children received a diagnosis of autism. This implies that if a child is simply slow in developing aspects of joint attention, this may be an indicator of language or general developmental delay. If however a child is severely impaired in SAM, an impairment in ToMM is an inevitable consequence, and this pattern carries a very high risk for autism.

ACKNOWLEDGEMENTS

The screening study reported here was supported by the MRC. We are grateful to Tony Cox, Gillian Baird, Auriol Drew, Kate Morgan, Natasha Nightingale, and Tony Charman – our colleagues on the screening study; and to Gabriel Segal and Peter Carruthers for valuable discussions.

NOTES

1 Its evolutionary history is also discussed by Baron-Cohen (1995a,b), and its neurological basis is discussed by Baron-Cohen and Ring (1994a).

2 The term 'triadic representation' is derived from Bakeman and Adamson's (1984) term 'triadic relation', which exists between two agents and a third object. They distinguish this from 'dyadic relations' which only involves a relation between two agents. Here, the focus is on *representing* these different kinds of social relations.

3 Povinelli and Eddy (1994) argue that simpler forms of gaze-monitoring may be possible without SAM (i.e., without any understanding of what I or you are *seeing*). Protodeclarative pointing, rather than gaze-monitoring, may therefore be the acid test of SAM. In humans, though, it is likely that both depend on SAM.

4 In Baron-Cohen (1994) a triadic representation was defined as *[Self-Relation-(Agent-Relation-Object)]*. We are grateful to Gabriel Segal for suggesting that

this representation should contain a proposition rather than simply an object, since infants appear to be able to represent propositions (Baillargeon, 1987).

5 Gabriel Segal and Peter Carruthers, our philosopher guides here, suggest the distinction we are groping for is that SAM's mental states can be thought of as *de re*, rather than *de dicto*. We are grateful to them for this suggestion, and return to this point later.

6 Note that others have argued this analysis of pretence is unnecessarily rich – e.g., Harris and Kavanaugh (1993), or Perner (1991a); but see Leslie (1994a) for a rejoinder to this criticism.

7 This is Leslie's term, which replaces his earlier term 'metarepresentations'. This change in terminology arose because of other authors also using the latter term (e.g., Pylyshyn, 1978; Perner, 1991a), but with a different definition. For the terminological wrangles, see Perner (1993) and Leslie and Roth (1993).

8 There remains an important question as to why, in normal development, some intentional terms (such as pretending) are understood before others (e.g., knowing) which in turn appear to be understood before yet others (e.g., believing). We do not address this issue here, but note it as an unresolved problem (see Leslie, 1994a, for one attempt at a resolution; or Harris, this volume, for another).

9 Another way of putting this, following Segal and Carruthers' suggestion, is that M-Representations represent mental state concepts that are *de dicto*, not just *de re*.

10 Lock and key hypotheses are of course widespread in biology: for example, DNA base pairs can only fit together in a specific way; antibodies can only fit on to the surface of specific antigens; certain hormones will trigger one mechanism to function but not another; etc.

11 We retain the terms *relation* within triadic representations, and *attitude* within M-Representations, to mark this distinction between the partially intentional nature of the former, and the fully intentional nature of the latter. However, in normal usage, these terms are obviously not mutually exclusive.

12 See Baron-Cohen, Allen, and Gillberg (1992), where the CHAT is fully described and its use in a genetic high-risk study of autism is reported.

13 In the CHAT, gaze monitoring involves following another person's pointing gesture and change of gaze direction. It could therefore more properly be called gaze-and-point-monitoring.