Can the 'Theory of Mind' Hypothesis Survive, Given Theoretical Insights Derived from the Study of Autism? A Response to Hacking and McGeer

Matthew Cull

Abstract

In this paper I agree with both Ian Hacking and Victoria McGeer that the 'Theory of Mind' theory (ToM) is fundamentally flawed. However, I find reasons to reject both of their critiques of ToM as incoherent and instead build upon certain parts of McGeer's work to develop my own rejection of ToM. I end by suggesting routes this rejection might take the philosophy of psychology down.

1 Introduction

The dominant paradigm within contemporary psychology suggests that humans have the capacity to understand the behaviour of others through the positing of theories of mind in the mode of a scientist at work. This theory in the philosophy of psychology is known as the 'Theory of Mind' hypothesis, or "Theory-Theory" (henceforth 'ToM'). Ian Hacking and Victoria McGeer have argued against this hypothesis for reasons derived from the study of Autistic individuals. In section two I will introduce ToM. In section three I will develop a reading of Hacking's critique of ToM, before showing how McGeer undermines this critique. In section four I will show that McGeer's own reasons for rejecting the 'Theory of Mind' hypothesis do not hold up to scrutiny. However, not to confirm ToM, in section five I shall draw on some other aspects of McGreer's thought to show that we should remain sceptical of ToM and will end by hinting at some of the possible implications of that rejection.[1]

2 The Theory of Mind Hypothesis

The origins of the 'Theory of Mind' hypothesis are generally considered to be found in Premack and Woodruff's *Does the Chimpanzee have a*

Kriterion – Journal of Philosophy (2014) 28: 45–54.
http://www.kriterion.at

Theory of Mind.[2] They define having a theory of mind as an individual "imput[ing] mental states to himself and to others".[3] This is considered 'theorylike' as one is positing unobservables (mental states) and making predictions about behaviour based on systematic ideas about those unobservables.[4] Developmental psychologists Gopnik and Metzoff expand upon this thought, arguing that a child initially has "theorylike structures for organizing the world, which he then modifies and reshapes in later theorizing".[5] Such theoretical structures include theories of mind, most importantly thinking of "themselves and others as sharing the same psychological states." Thus, according to the ToM hypothesis, not only do we understand others through the positing of a theory of mind, we also only come to understand ourselves through a similar hypothesis.[6]

Obviously, there is a vast literature on the 'theory of mind'. I take myself, in this essay, to be responding to ToM as formulated by the likes of Gopnik and Metzoff, but it remains an open question as to the arguments I raise here apply to other specific formulations of the theory of mind. Henceforth, whenever I make reference to 'the ToM' or 'the ToM theorist', I am making reference to the version of ToM explicated above and in the writing of Hacking and McGeer.

Similarly, later in this essay, I will make reference to autism and results from the study of autism as a part of challenge posed to this form of ToM, yet to pose a monolithic, singular conception of autism would fail to do justice to the wide variety of autistic experiences. It is often said that 'if you know one person with autism, then you know one person with autism' – the point being that the spectrum of autistic individuals is so varied that making generalisations about them or their experience is impossible. Therefore, in referring to 'the autistic person' later in this paper, let it be known that I will be referring to specific aspects of many (but not all) autistic people's experiences.

3 Hacking's Critique of the Theory of Mind Hypothesis

Ian Hacking criticises the view of the mind described by ToM, but it is difficult to discern a direct line of argument against it in any of his writings. Hacking writes that ToM posits a faculty of mind "for attributing mental states".[7] But, as he points out, from the point of view of ToM, it is difficult to see whether the act of attribution is inferential or simply read "right off" the action one observes.[8] To demonstrate this latter point, Hacking invokes Köhler's phenomena – examples of where we directly observe and therefore do not infer mental states. Such examples

are so commonplace as to be banal – I, sat in a library, can see who is focusing and who is relaxing, simply by looking at them. I do not need to make inferential judgments of the form "they are surrounded by books, their brow is furrowed and they are hunched over their work, therefore they are focusing hard" in order to determine a person's mental states. Of course, as I have just demonstrated, it is possible to go through such an inferential process, and indeed, were I asked *why* I thought that person was focusing hard, I might justify myself in such a manner. [9] ToM, being grounded in a picture of a module in the mind that posits theories, testing and refining them,[10] cannot, on this construction of Hacking's account, deal with Köhler's phenomena, as beliefs about the mental states of others arising from Köhler's phenomena are non-inferential.

Autism plays into Hacking's criticism of ToM thus – according to him, autistic people do not experience Köhler's phenomena.[11] On the ToM account, Autistic individuals are simply considered "mind-blind," failing to have a good theory of mind due to a faulty ToM module.[12] Yet it would appear to imply that those without a good theory of mind, or ToM, will fail to make good *inferential* judgments about the mental states of others. This is contrary to the experience of autistic individuals, who seem fine with inferential judgments about the mental states of others, but fail to experience Köhler's phenomena. Indeed, having a 'poor theory of mind' does not explain the difference between failing to understand the behaviour of another inferentially and failing to understand due to a lack of Köhler's phenomena well. It appears, on this reconstruction of Hacking, that ToM has insufficient explanatory power to account for autistic experience.

I am, however, inclined to agree with Victoria McGeer, that "ToM advocates do not see their approach as requiring explicit inferential processes".[13] In hypothesising a sub-personal module that deals with a ToM, the 'theorising' need not be done consciously. Thus Hacking (or at least our presentation of his thought) has it wrong; the ToM module does not deal with explicit inferential reasoning about other behaviour, but rather, it deals with Köhler's phenomena. To say one has a faulty ToM module is to say one has a lack of, or impaired access to, Köhler's phenomena. ToM is therefore perfectly explanatorily valid with respect to autism.

4 McGeer's Critique of the Theory of Mind Hypothesis

McGeer however, does agree with Hacking that ToM theory fails in the face of evidence drawn from research on Autism. She claims the "ToM deficit hypothesis implies that the autistic inability to experience Köhler's phenomena is a disability unique to them, consequent on damage or dysfunction to their mind-reading system." But this is to ignore, she argues, the failure of neurotypical individuals to experience Köhler's phenomena when interacting with Autistic individuals.[14] Imagine, for a moment, a conversation between a high-functioning autistic person and a neurotypical person. If the ToM hypothesis holds, then the neurotypical (with a working ToM module) ought to be able to experience Köhler's phenomena when observing the autistic persons' behaviour, 'reading off' the mental states of the autistic person. Meanwhile, the autistic person (with a faulty ToM module) would be unable to experience Köhler's phenomena in observing the neurotypical person's behaviour, failing to 'read off' the mental states of the neurotypical person. Thus the ToM hypothesis predicts an asymmetry in Köhler's phenomena in autistic-neurotypical interactions, as the neurotypical would experience Köhler's phenomena whilst the autistic person would not. If there is a deficit only in the autistic mind's ToM module, as the ToM hypothesis claims, there would be no reason to suspect that the neurotypical mind fails to perceive Köhler's phenomena in autistic individuals. Yet McGeer claims that this is not the case – given that there is actually a symmetrical lack of Köhler's phenomena in autistic-neurotypical interactions. She thinks that neurotypical individuals *do not* actually experience Köhler's phenomena when observing the behaviour of autistic individuals. Given that the ToM hypothesis predicts asymmetry, ToM must be false and thus rejected.

To put this another way, the ToM hypothesis predicts that a neurotypical person will experience Köhler's phenomena when observing the actions of an autistic person – as given neurotypical individuals have a working ToM module, that module will impute mental states to the autistic person. McGeer, however, holds that this simply does not occur – therefore we ought to reject the ToM hypothesis as a model of the mind.

In place of the ToM, McGeer, again influenced by Hacking, posits that being unable to experience Köhler's phenomena is due to a failure, not in a ToM module of the mind, but rather in one's knowledge of a 'form of life'. This, rather Wittgensteinian thought (indeed Hacking himself derives his thought directly from the *Philosophical Investiga-*

tions and *Remarks on the Philosophy of Psychology*) expresses the idea that one must understand the norms and rules of a language-game another person is playing in order to understand that person's expressions, as those expressions will only make sense in the context of that language game.[15] This is, in some sense, what Wittgenstein meant when he claimed "if a lion could talk, we could not understand him," for we do not share a 'form of life' with any lion.[16] McGeer uses the idea of 'being a friend' in order to explain this thought. Being a friend consists in not letting others down, not talking behind their back and even laughing at their jokes – such norms of friendship are understood by anyone skilled at being a good friend. As McGeer puts it, "Anyone who's skilled in the practice of being a friend understands these things, and regulates himself accordingly, depending on the level of friendship he hopes to sustain in a particular relationship. Equally, anyone who's skilled in the practice of being a friend, recognizes when others are, through their actions and expressions, either trying to adhere to the norms of friendship, only paying lip service to these norms, or actively showing a disinterest in observing them."[17] What McGeer thinks of as the symmetrical nature of lack of Köhler's phenomena in autistic-neurotypical interaction is thus explained by autistic and neurotypical individuals failing to share a common 'form of life' or language-game, whereby they might understand one another. Autistic individuals are "simply not skilled in the myriad practices that constitute our shared form of life".[18] McGeer seems to want to stay relatively neutral on the causes of this lack of skill, but thinks that a complex developmental story will need to be told that takes into account both endogenous and exogenous factors that prevent autistic individuals from joining the complex social language games that make up neurotypical society.[19]

However, an experiment McGeer herself brings up casts doubt upon her conclusions. First performed in the 1940s by Heider and Simmel, individuals are asked to describe a film consisting of moving shapes. Neurotypical individuals tend, overwhelmingly, to describe the shapes in anthropomorphic terms, describing even the movement of two triangles as "a stirring little drama involving two friends". In contrast, when high-functioning individuals participated in the experiment, such narratives were notably absent, along with (most crucially) language referring to mental states. The lack of mental state terminology in describing the film, suggests McGeer, is attributable to autistic individuals not experiencing Köhler's phenomena, whilst neurotypical individuals' use of mental state vocabulary implies that they experience Köhler's phenom-

ena.[20] This strikes me as odd. Neurotypical individuals fail to experience Köhler's phenomena when dealing with autistic individuals, according to McGeer, yet are somehow able to experience Köhler's phenomena when looking at abstract shapes. Further, in all of McGeer's work on the subject that I have been able to access[21][22][23] not once is there an empirical study to back up the former assertion. Given this, her argument from the symmetrical lack of Köhler's phenomena against ToM fails, as I claim there is simply no such symmetry.

McGeer might argue that neurotypicals fail to experience Köhler's phenomena by responding that there is something especially difficult about attributing mental states to autistic individuals, but given neurotypical individuals appear capable of imputing mental states to abstract objects such as triangles, this seems implausible. Alternatively, she might argue that it is particularly difficult to attribute the *correct* mental states to autistic individuals in a way that does not apply to triangles. Certainly, there are no standards for correctness in our description of a triangle's mental states as it moves across a screen and clearly there are standards of correctness when describing another person. However, the question is not whether the neurotypical individual's Köhler's phenomena correctly line up with an autistic person's mental states, but rather, whether or not neurotypical individuals experience Köhler's phenomena in such circumstances at all. Of course, armchair theorising will not settle this debate – empirical study is the only way to find out if neurotypical individuals do experience Köhler's phenomena when interacting with autistic individuals. However, given the arguments presented thus far, we have no reason to believe that neurotypical individuals fail to experience Köhler's phenomena in their interaction with autistic individuals and therefore have no reason (at least from McGeer's main argument) to reject ToM.

One might claim that McGreer was wrong to characterise neurotypical responses to Heider and Simmel's experiment as those of Köhler's phenomena and therefore a lack of Köhler's phenomena when interacting with autistic individuals becomes more plausible for the neurotypical. In such circumstances, one could perhaps argue that neurotypicals in the experiment are using inferential reasoning to attribute mental states to triangles. This however, fails to explain why autistic individuals usually do not attribute mental states to the abstract shapes in the experiment. If the neurotypical person is able to inferentially reason to mental states for triangles, it seems odd that the autistic person cannot.

5 Autistic Autobiography as Critique of the Theory of Mind Hypothesis

There is, however, more to be said for a rejection of the ToM hypothesis. McGeer touches on this when she reminds us that ToM theorists must reject, or at least regard with suspicion, autistic autobiography.[24] Given the ToM theorist believes that the ToM module is essential for the swift attribution of mental states through Köhler's phenomena, even to oneself, the ToM theorist appears committed to the idea that those with a faulty ToM module will fail to accurately describe their own experience.

However, it seems strange to talk of being unable to describe one's own phenomenal experience correctly. Of course, mistakes of memory are quite ordinary, but the ToM theorist is not pointing to a flaw in the autistic person's memory. Instead, they are pointing to the manner in which the autistic person describes the way they experience the world (devoid of Köhler's phenomena) and claiming that this narrative of their phenomenal experience is somehow incorrect or invalid. Their position is akin to saying were a neurotypical person were to take the autistic person's place, the narrative of that person's phenomenal experience would be more accurate. Of course this is nonsense – the accuracy of the description of one's phenomenal experience does not depend on the content of one's phenomenal experience, not least whether that content matches up to the world. I may be incorrectly interpreting the world by thinking that there is a cat in front of me when in fact there is a dog – however, a narration describing my phenomenal experience at some later time would be wrong were I not to claim that I once thought there was a cat before me.

If the ToM hypothesis in conjunction with evidence from autistic individuals leads to such incoherent conclusions, I am tempted to reject it entirely.[25] This itself is a result in the philosophy of psychology – but let us examine its wider implications briefly.[26] ToM holds that Köhler's phenomena are only experienced by those with a working ToM module. The ToM module posits mental states and predicts the behaviour of others in the manner of a scientific theorist. If this hypothesis does not hold, we need a new theory of Köhler's phenomena, such as the above language-game theory explicated by McGeer, albeit modified to take into account the Köhler's phenomena experienced by neurotypical individuals when interacting with autistic individuals.

I should like to conclude with some remarks on what a rejection of ToM might mean for the philosophy of psychology more widely, if we are right to think that the remarks made about ToM as understood

here apply to ToM broadly construed. I suggest that the philosophy of psychology would become more focused on social psychology, linguistics and even sociology, above neurobiology and cognitive science. We further might see an increased appreciation of the historically situated nature of the mind. No longer would we think of a biologically determined ToM module which governs the Köhler's phenomena we experience – instead we find contingent and mutable language-games, the shared play of which generate Köhler's phenomena. Philosophy of psychology, if we are right to be sceptical of ToM in the light of autism, may learn to be a very different field of inquiry.

Notes

1 I would like to thank Professor David Bakhurst and Ryan McInerney for inspiring this paper and the helpful comments they offered over the course of its development, along with comments from anonymous reviewers at the Kriterion – Journal of Philosophy.

2 [9]

3 The imputation of states, such as belief and anger, to oneself and others, does not, for Woodruff and Premack, have to be completely accurate for there to be a theory of mind present. Nor need that theory impute all of the states that others impute.

4 [9, p.515]

5 [2, p.126]

6 Ibid., p.133.

7 [5, p.54]

8 [4, p.1470]

9 Another justification may be of the form "Just *look* at them!" This form of explanation being reliant on both myself and my interlocutor being able to experience Köhler's phenomena.

10 [2, p.11]

11 [4, p.1471]

12 [3, p.53]

13 [8, p.523]

14 Ibid., p.524.

15 Ibid., pp.524–5.

16 [10, p.223]

17 Ibid., p.525.

18 [8, p.525]

19 Ibid., pp.525—6

20 [8, p.522]

21 [6]

22 [7]

23 [8, pp.517–530]

24 [8, p.527]

25 In addition to the largely logical and linguistic discussion above, there are political arguments against ToM to be made here, regarding the valuation of 'difference' in society, and the inclusion of a variety of positions and knowledges. The autistic narrative might be considered emancipatory, the autistic mind not deficient, but different and valuable for that difference. Further, the reducing the autistic person to the level of a chimpanzee (let us not forget the origins of ToM) seems to preclude treatment of such individuals as persons equal to neurotypical individuals. Such discussion lies beyond the scope of this essay however.

26 There is an interesting question as to whether a rejection of the ToM favors the Russian tradition (that 'school' developed by the likes of Evald Ilyenkov and Felix Mikhailov – see, for an introduction, David Bakhurst's "*Consciousness and Revolution in Soviet Philosophy, From the Bolsheviks to Evald Ilyenkov*": [1]) in the philosophy of mind. Certainly, if we accept McGeer's suggestion, we begin to lean towards an increased emphasis on the social as constitutive of mind, simply because of the social nature of language games. Hacking's influence here should not be underestimated.

Matthew Cull
University of St Andrews
10 Delavale Road,
Winchcombe,
Cheltenham,
Cloucestershire,
GL54 5HN,
United Kingdom

<mcull117@gmail.com>

References

[1] David Bakhurst. *Consciousness and Revolution in Soviet Philosophy, From the Bolsheviks to Evald Ilyenkov*. Cambridge University Press, Cambridge, 1991.

[2] Alison Gopnik and Andrew N. Meltzoff. *Words Thoughts and Theories*. The MIT Press, Cambridge, MA, 1997.

[3] Alison Gopnik, Andrew N. Meltzoff, and Patricia K. Kuhl. *The Scientist in the Crib: Minds, Brains and how Children Learn*. William Morrow and Company Inc., New York, 1999.

[4] Ian Hacking. Autistic Autobiography. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1522):pp.1467–1473, 2009.

[5] Ian Hacking. Humans, Aliens and Autism. *Daedalus*, 138(3):pp.44–59, 2009.

[6] Victoria McGeer. Psycho-practice, Psycho-theory and the Contrastive Case of Autism: how practices of mind become second-nature. *Journal of Consciousness Studies*, 8(5–7):pp.109–132, 2002.

[7] Victoria McGeer. Autistic Self-Awareness. *Philosophy, Psychiatry and Psychology*, 11(3):pp.235–251, 2004.

[8] Victoria McGeer. The Thought and Talk of Individuals with Autism: Reflections on Ian Hacking. *Metaphilosophy*, 40(3–4):pp.517–530, 2009.

[9] David Premack and Guy Woodruff. Does the Chimpanzee have a Theory of Mind? *Behavioural and Brain Sciences*, 1(4):pp.515–526, 1978.

[10] Ludwig Wittgenstein. *Philosophical Investigations*. Basil Blackwell, Oxford, Trans. Anscombe G.E.M. edition, 1986.