OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1037/rev0000532

Document Version
Publisher's PDF, also known as Version of record

Link to publication record in King's Research Portal

Citation for published version (APA):
Long, E. L., Catmur, C., & Bird, G. (in press). The theory of mind hypothesis of autism: A critical evaluation of the status quo. *Psychological Review*. https://doi.org/10.1037/rev0000532

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 07. Oct. 2026

Psychological Review

The Theory of Mind Hypothesis of Autism: A Critical Evaluation of the Status Quo

Emily L. Long, Caroline Catmur, and Geoffrey Bird

Online First Publication, January 9, 2025. https://dx.doi.org/10.1037/rev0000532

CITATION

Long, E. L., Catmur, C., & Bird, G. (2025). The theory of mind hypothesis of autism: A critical evaluation of the status quo. *Psychological Review*. Advance online publication. https://dx.doi.org/10.1037/rev0000532

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2025 The Author(s)
ISSN: 0033-295X

https://doi.org/10.1037/rev0000532

THEORETICAL NOTE

The Theory of Mind Hypothesis of Autism: A Critical Evaluation of the Status Quo

Emily L. Long[1], Caroline Catmur[2], and Geoffrey Bird[1, 3]

[1] Department of Experimental Psychology, University of Oxford

[2] Department of Psychology, King's College London

[3] Centre for Research in Autism and Education, Institute of Education, University College London

The theory of mind (ToM) hypothesis of autism is the idea that difficulties inferring the mental states of others may explain social communication difficulties in autism. In the present article, we critically evaluate existing theoretical accounts, concluding that none provides a sufficient explanation of ToM in autism. We then evaluate existing tests of ToM, identifying problems that limit the validity of the conclusions that may be drawn from them. Finally, as an example of how the identified issues may be resolved, we describe work developing a psychological account of ToM (the Mind-space framework) and a new test of ToM accuracy (the Interview Task).

Keywords: theory of mind, mentalizing, autism, neurodiversity, mind space

The theory of mind (ToM) hypothesis of autism is perhaps the most influential and long-standing cognitive theory attempting to explain the features of autism spectrum disorder (henceforth "autism"). Various iterations of this hypothesis posit that a specific cognitive deficit in ToM characterizes autism and that many autistic symptoms can be attributed to this deficit (Frith et al., 1991; Leslie & Frith, 1987). ToM has been classically defined as the ability to represent the mental states of oneself and others (Premack & Woodruff, 1978), and much existing work has concluded that there are clear differences in this ability in autistic individuals relative to neurotypical individuals (Baron-Cohen et al., 1985; Dziobek et al., 2006; F. G. Happé, 1994; for meta-analyses, see, e.g., Gao et al., 2023; Yirmiya et al., 1998).

The longevity and widespread acceptance of the ToM hypothesis is largely explained by the success it is thought to have in explaining the difficulties with social communication and social interaction seen in autism. For example, there is a substantial body of evidence suggesting impairments in autistic individuals' understanding of pragmatic language and sarcasm, as well as in their recognition of social faux pas (Baron-Cohen et al., 1999; Frith et al., 1994; Reindal et al., 2023; Thiébaut et al., 2016), which may be explained by differences in ToM ability. When a person's utterance is ambiguous, as in the case of sarcasm and some aspects of pragmatic language (e.g., metaphor), it may be best understood by appealing to their mental state. For example, if a friend exclaims "Oh, great!" after dropping a plate of food, a correct understanding of this utterance cannot be reached solely by processing the verbal content (which would lead one to believe that the friend considers dropping the food to be a positive event). Instead, one may correctly interpret the sarcasm by inferring that the friend believes dropping the food to be bad and that therefore they intend to communicate their frustration, rather than their pleasure, at this happening. In this case, an inability to represent (or accurately infer) a conversation partner's mental states would lead to a disadvantage in understanding their utterance and thus in responding appropriately. Indeed, there is a substantial body of work claiming a link between ToM and both social functioning and pragmatic language understanding in individuals with and without autism (Andrés-Roqueta & Katsos, 2017; Bosco et al., 2018; Peterson et al., 2016; Tager-Flusberg, 2003).

However, if one wishes to understand autistic cognition or to develop effective interventions to support autistic individuals, it

Elena L. Grigorenko served as action editor.

Geoffrey Bird https://orcid.org/0000-0002-2310-0202

Geoffrey Bird is supported by the Baily Thomas Charitable Fund and the John Templeton Foundation. Caroline Catmur is supported by the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. The authors are beyond grateful to Lucy Foulkes, Francesca Happé, and Cecilia Heyes for reading and suggesting edits on earlier drafts of this article. It is truly wonderful to be able to call on such lovely, intelligent people for their advice.

Open Access funding provided by University of Oxford: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; https://creativecommons.org/licenses/by/4.0). This license permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Emily L. Long played a lead role in writing–original draft and an equal role in conceptualization and writing–review and editing. Caroline Catmur played a supporting role in conceptualization and an equal role in writing–review and editing. Geoffrey Bird played an equal role in conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to Geoffrey Bird, Department of Experimental Psychology, University of Oxford, Anna Watts Building, Woodstock Road, Oxford OX2 6GG, United Kingdom. Email: geoff.bird@psy.ox.ac.uk or geoff.bird@ucl.ac.uk

would be valuable for ToM differences in autism not only to explain social communication difficulties in autism but to themselves be explained. To achieve this, one needs a psychological model of the processes underlying ToM, specific hypotheses about how they may differ in autism, and valid and reliable measures for testing these hypotheses. In this article, we critically evaluate the extent to which existing work on the ToM hypothesis of autism might be said to meet these criteria.

We begin by describing two key conceptual challenges that arise in this area, highlighting the need to clearly define ToM ability before developing cognitive theories of ToM in autism. Then, we trace the development of psychological theory from the original proposal of the ToM hypothesis to the present day, concluding that no proposal offered thus far provides a sufficient explanation of hypothesized ToM difficulties in autism. We close our assessment of the status quo by evaluating existing tests of ToM ability, identifying limitations that may mean that existing tests are unable to test any psychological model or even conclusively support the notion that autistic individuals truly show deficits in ToM relative to neurotypical individuals. Finally, we describe our own attempts to resolve the identified issues as one example of how the ToM hypothesis may be reconceptualized.

Defining the Conceptual Space of ToM in Autism

In this section, we outline how the lack of shared understanding in two key aspects of the definition of ToM contributes to difficulties in adequately specifying psychological models of ToM in autism. Specifically, we explore the definition of a "mental state" and consider whether ToM might be better conceived as a process of inference, not (only) of representation. By evaluating these issues, we begin to examine what may be required of a model of ToM in autism. Several of the concepts we discuss throughout this section are summarized in Table 1.

Defining a "Mental State"

Mental states are classically defined as propositional attitudes (Leslie, 1987; Premack & Woodruff, 1978). A propositional attitude is an agent's mental relationship to a specific proposition, where a proposition is considered a declarative statement about the world. For example, "the sky is blue" is a proposition about the color of the sky, which may or may not be true in the real world. Agents may have different attitudes to the truth of different propositions (e.g., one might *believe* that the sky is blue but *wish* that the sky is pink) or even have multiple attitudes toward the truth of the same proposition (e.g., one might *intend* that they appear friendly but concurrently *disbelieve* that they appear friendly). Similarly, different agents may have different propositional attitudes: One member of a household might believe that *the chocolate is in the cupboard*, but their housemate might not believe that to be true because they instead believe that *the chocolate is in the fridge*. Thus, a proposition is a statement about the physical state of the world (which may or may not be true), while a propositional attitude is a statement about a mental attitude toward a proposition, not about the physical state of

Table 1
Summaries of Key Concepts

Concept	Summary
Theory of mind (ToM)	ToM is classically defined as the ability to attribute mental states to self and others to explain and predict behavior (Premack & Woodruff, 1978). In practice, however, many tests assume that the capacity for mental state representation is present in all participants and instead test the accuracy (or typicality) of mental state inferences (Conway & Bird, 2018; Pisani et al., 2021). ToM may therefore be better defined as the inference of mental states, which entails the ability to represent such states. Levels of ToM ability, then, may be considered as the accuracy of such inferences or the propensity to make them.
Mental state	A mental state is defined as a propositional attitude (Leslie, 1987; Premack & Woodruff, 1978). However, many tests of ToM include emotional states, or states of mind (e.g., tiredness, drunkenness) as mental states. It is not clear that the psychological requirements for inferring propositional attitudes, emotional states, and states of mind are the same, and so a clear and precise definition of what constitutes a mental state is required when considering ToM in autism. Here, however, we follow the classical definition, considering mental states solely as propositional attitudes.
Propositional attitude	A proposition is a statement that can be true or false. For example, "the sky is blue," "Lucy has a chocolate bar," or "I will make a good impression in the interview." A propositional attitude, then, is an agent's attitude to the truth of a particular proposition, and these attitudes are typically beliefs, desires, or intentions. For example, "Sarah believes that the sky is blue," "Lucy wishes that she has a chocolate bar," and "I intend that I will make a good impression in the interview."
Theory of mind hypothesis	The theory of mind hypothesis of autism is the suggestion that difficulties in understanding mental states (particularly the mental states of other agents) explain many of the social communication difficulties observed in autism (Frith et al., 1991).
Representation	Some theories seeking to explain the ToM hypothesis of autism, most prominently the theory of mind module account, suggest that autistic individuals are unable to *represent* mental states. An inability to represent mental states would be an inability to hold in mind, in any form, a propositional attitude with the realization that propositional attitudes can be decoupled from reality. The capacity to represent is all-or-none (one either can hold something in mind in this form or they cannot), and therefore these theories necessarily suggest that autistic individuals are incapable of performing ToM in any form. If this is the case, autistic individuals should perform at chance levels at any true ToM test that cannot be performed through other means (i.e., without appealing to mental states).
Inference	Although many accounts define ToM only as the capacity to represent mental states, its original definition suggested that a process of inference is required for representation to occur (Premack & Woodruff, 1978), and the majority of ToM tests assume that all participants can represent mental states. Instead, tests focus on the *inference* of mental states: the process of making use of available information to determine the most likely content of a propositional attitude. The results of these tests suggest that differences in ToM in autism do not constitute an inability to represent mental states but instead indicate differences in the process of *inferring* another's mental state.

the world. A propositional attitude can therefore be described as "decoupled" from physical reality—its validity cannot be determined by reference to the physical world.

Mental states, as defined above, are distinct from, but sometimes related to, other types of mental events such as emotions. For example, the process of evaluating whether someone desired a particular item might be easier if one can interpret their emotional expression upon receiving it. Similarly, the process of inferring whether an individual holds a correct belief about the location of an item might involve tracking their visual perspective to understand what information is available to them. The processing of (some of) these related mental events, as well as other states that the mind may occupy (e.g., tiredness, thoughtfulness, or drunkenness), is sometimes considered to constitute ToM (Baron-Cohen et al., 2001; Dziobek et al., 2006; Samson et al., 2010; Saxe & Houlihan, 2017; Senju, 2012; Tamir & Thornton, 2018) and sometimes not (Leslie & Frith, 1988; Oakley et al., 2016). It seems that there is little shared understanding about what a "mental state" is and therefore, in turn, about what it means to "represent a mental state" and thus to perform ToM.

It is not the case that the classical definition of a mental state is inherently correct: If evidence suggested that precisely the same psychological processes underpin the understanding of propositional attitudes as the understanding of emotional states, then the distinction between these two constructs might be arbitrary and ultimately irrelevant to psychological modeling of these processes. However, if it is not believed (and explicitly stated) that propositional attitude inference and emotional state inference are identical processes, it would be inappropriate to test a theory that purports to explain the inference of propositional attitudes using a test of emotional inference. We will explore, in our evaluation of existing tests of ToM, inconsistencies between the understanding of ToM in psychological theory and its operationalization in testing, but it should be noted here that such issues can only be avoided through both careful test design *and* the development of well-specified theory.

For the remainder of this article, we use the term "mental states" to refer specifically to propositional attitudes.

An Ability "to Represent" or an Ability "to Infer"

ToM is frequently defined as the ability to *represent* the mental states of oneself and others, following Premack and Woodruff (1978). Premack and Woodruff sought to test whether chimpanzees could represent mental states as belonging to themselves and to others but also recognized that, because mental states are not directly observable, such states need to be inferred based on observable cues. Both representation (holding mental states in mind as propositional attitudes and thus decoupled from physical reality) and inference (determining the content of the mental state) are necessary parts of ToM: The representation of a mental state as belonging to an agent may only be beneficial for social interaction if the agent truly holds that mental state. Any theory of ToM in autism should specify which of these abilities is thought to differ between autistic and neurotypical individuals because the implications of a representational impairment differ substantially to the implications of an inferential impairment.

The *capacity to represent* mental states/propositional attitudes is necessarily binary. One either has the capability to hold propositions in a form decoupled from reality or one does not. One individual may be slower to realize that mental state representation is necessary than another or have a smaller working memory and so can hold fewer mental states in mind at any given moment (or less complex propositional content), but if representing mental states means the ability to hold mental states in mind as propositional attitudes with the potential for their propositional content to be decoupled from physical reality, then these are not differences in the ability *to represent* mental states. To be clear, it is obviously the case that some propositions are harder to represent than others—"A > B, C > D, C > A, but B is > D" is harder to represent than "A > B." However, if I add "Sarah believes …" before each of these propositions, the difference in ease between representing each as a proposition and as a propositional attitude does not vary between them. One either understands that propositional attitudes are decoupled from reality or one does not.

Given this, it is not clear how, if ToM is defined solely as the ability to represent mental states, some mental states (and therefore some tests) could be more difficult than others in terms of mental state representation. Instead, differences in performance between different tests of mental state representation may be better explained by the demands they place on other abilities, such as language or working memory (Arslan et al., 2017; Filip et al., 2023; Hale & Tager-Flusberg, 2003), or by the conceptual knowledge they require, for example, what must be known about minds (Conway & Bird, 2018), rather than by the difficulty of the necessary mental state representations. In other words, the difference in the difficulty associated with representation of certain mental states is associated with the content of the proposition, not with representing that proposition as a mental state. Because of this, the reliable success of some autistic participants across tests of mental state representation (Frith et al., 1994; Gernsbacher & Yergeau, 2019; F. G. Happé, 1994) is problematic for accounts that hypothesize impaired mental state representation in autism, regardless of whether this evidence is found in every test or every autistic individual. We will return to this notion when evaluating existing theoretical accounts of ToM in autism.

As already noted, another necessary component of ToM is the ability to *infer* mental states, to make use of the information available to hypothesize about the content of an agent's mental state. Mental state inference could be considered as akin to physical inference, in which information is gathered and combined to understand the physical structure of the world and the forces within it in a manner that facilitates predictions about causal events (Fischer et al., 2016). As in physical inference, mental state inferences can vary in difficulty, and individuals can vary in their ability to make inferences.

The difficulty of an inference might vary due to the amount of information present in the environment (e.g., the visibility of part of a physical scene or the detail with which a person describes their situation) or the amount of prior knowledge one possesses (e.g., the novelty of a certain material or our familiarity with the person whose mental state is being inferred). Additionally, inferences might be more difficult if they require consideration of a greater number of variables and cannot be simplified by reliance upon a limited subset of diagnostic variables. For example, the trajectory of a ball may be harder to infer when it is colliding with another moving object than when it is colliding with a simple barrier. Similarly, predicting one agent's belief about another's intentions may be difficult because the belief is a product of both agents' minds (i.e., of the latter's intentions and the former's interpretation of them) but may be easier if the inference can be based on a reduced set of highly diagnostic

information (e.g., if the intention being inferred is the intention to run away from a chasing bear).

Differences may not only appear in the difficulty of inference problems but also in individual inference ability. Different individuals might vary in the accuracy of their inferences (e.g., how accurately one can predict where a ball will come to rest or the extent of a person's intention to offend with an ambiguous remark) or their propensity to make them (i.e., whether they routinely consider the trajectories of objects or people's intentions), but they are unlikely to vary in their ability to represent physical or mental states (i.e., whether they *could* conceive of a possible trajectory of the ball or the possible extent of the person's desire to offend).

To further illustrate, consider the design of functional magnetic resonance imaging (fMRI) studies attempting to identify the neural correlates of ToM. These studies often contrast an experimental condition requiring representation of mental states with a control condition, which requires representation of a nonmentalistic property (see, e.g., Saxe & Kanwisher, 2003). In such studies, provided the control condition was matched in all other features, any neural activity unique to the experimental condition would relate to the representation of mental states. However, any individual demonstrating this unique blood oxygenation level-dependent (BOLD) response may be completely incapable of making *accurate* mental state inferences despite being able to represent mental states. When asked to infer what someone intends, believes, or knows, they may consistently give the wrong answer, yet they have a full understanding of what it means to intend, believe, or know and the ability to represent mental states using unique neural (and possibly psychological) systems. Differences in the BOLD response itself could be indicative of differences in the representation of mental states, although, in practice, fMRI data are difficult to interpret in the context of ToM in autism, a problem to which we will return when examining existing tests.

By contrast, behavioral tests of ToM often purport to measure the accuracy of ToM inferences, a claim which we will evaluate later, but it should be noted here that although individual incorrect answers in these tests are ambiguous in terms of whether they indicate a deficit of representation or of inference, anything other than chance performance indicates preserved representational ability. This is because, given that representation is all or nothing, an individual with a deficit in representation could not interpret any behavior in terms of mental states or even truly understand the mental states detailed in the possible answers of multiple-choice tests. As such, they would be expected to respond randomly, performing at chance accuracy, if the test is sufficiently rigorous such that it cannot be passed using nonmentalistic means.

However, both neurotypical and autistic individuals consistently perform at above chance levels in tests of ToM accuracy (Dziobek et al., 2006; F. G. Happé, 1994; S. White et al., 2009), and indeed "advanced" tests were developed to resolve concerns that binary, all-or-nothing tests (i.e., tests of representation) could not detect group differences in ToM (Baron-Cohen, 1989; Frith et al., 1991; F. G. Happé, 1994). This suggests that errors in ToM tasks are typically errors of inference: Individuals get the answer wrong because, even though they are perfectly capable of representing mental states, they have attributed the wrong mental state to the agent. An account that argues that autism involves an inability to represent mental states, then, does not align with existing evidence, and a satisfactory cognitive account of the ToM hypothesis may therefore require consideration of the process of inference.

What Is Required of a Theoretical Model of ToM in Autism?

In this section, we have outlined two key conceptual issues in the understanding of ToM: We have highlighted the lack of agreement between ToM researchers regarding what exactly constitutes a mental state and demonstrated that the common definition of ToM as "the ability to represent mental states" does not provide a complete description of the processes required for mental state understanding. As we have noted, the notion of ToM as representation alone is not in line either with existing evidence, which indicates that the ability to represent mental states is (near) universal in human adults with sufficient intelligence, or with currently used tests (which were developed to increase the sensitivity and rigor of ToM measurement).

Consideration of these issues provides some insights into what might be expected of a satisfactory theoretical model of ToM in autism. Specifically, these lines of thinking suggest that a satisfactory model of ToM, one that can truly be said to explain group differences, is likely to offer suggestions as to how information might be processed in the inference of propositional attitudes (or some other clearly specified set of mental events) and how the operation of the necessary processes might differ in ways that result in atypical ToM.

Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism

Having outlined conceptual considerations to be made when specifying a model of ToM, we will now evaluate existing proposals, considering whether they provide sufficient explanations of ToM differences in autism, as well as their compatibility with existing empirical evidence. To do so, we will trace the development of psychological theorizing about the ToM hypothesis of autism from its original proposal, as in the theory of mind mechanism or module (ToMM) account (Frith et al., 1991; Leslie, 1987; Leslie & Frith, 1987), to the present day. As shown in Figure 1, theoretical developments have occurred along two primary paths. Along the first, theories increasingly tolerate the (ostensible) existence of (some) ToM ability in autism. Along the second, theories increasingly argue that ToM impairments in autism arise from domain-specific processing deficits.

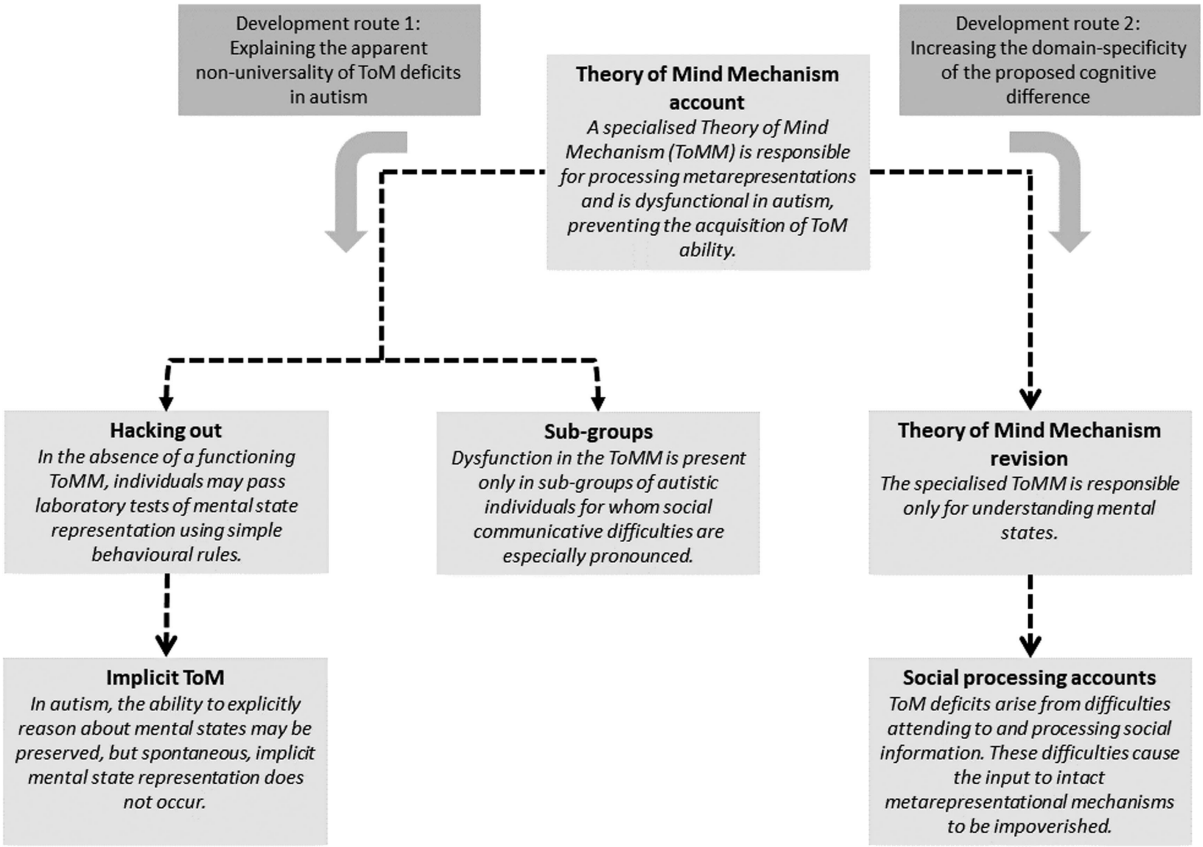
Classical Theory: The ToMM

Traditionally, it has been suggested that ToM is subserved by an innate neural mechanism, the ToMM (Leslie & Frith, 1987, 1990; Leslie & Thaiss, 1992). This mechanism allows an individual to understand and predict the behavior of a conspecific by representing their mental states, here defined as propositional attitudes. Importantly, the ToMM was said to be responsible for processing all metarepresentations (Leslie, 1994), not solely others' mental states.

In this body of literature (Leslie & Frith, 1987, 1990; Leslie & Thaiss, 1992), a metarepresentation is defined in terms of what the (meta-)represented proposition entails about the real world. According to Leslie (1987), whereas a primary representation is "transparent," meaning that the proposition it contains is considered to be a true representation of some aspect of the world, a metarepresentation is "opaque," meaning that its propositional content does not suggest anything about the real world. For example, the primary representation "the sky is blue" implies that the sky truly is blue. By contrast, the

Figure 1

A Diagram Summarizing the Development of Existing Theoretical Proposals Explaining the ToM Hypothesis



Note. ToM = theory of mind.

metarepresentation "Sally believes that the sky is blue" only entails that this is Sally's belief and thus leaves the possibility that the proposition "the sky is blue" is false. As such, holding the lower order proposition within the higher order metarepresentational context means that the representation no longer implies that the sky is blue.

Representing and processing a metarepresentation is said to require processing the opacity of the representation, i.e., that it does not entail that its propositional content is true ("decoupling"; Leslie, 1987). It should be noted that by defining metarepresentation in this manner, the ToMM account may be open to claims of circularity. To hold a metarepresentation (i.e., to hold a representation in a form that allows decoupling), the neurocognitive system (and subsequently the conscious human) must recognize that the proposition should be held in the metarepresentational context (i.e., that it may need decoupling). Problematically, this process of recognition itself requires that the proposition not be taken to entail facts about the world and thus itself requires metarepresentation.

Frith et al. (1991) argued that heterogeneous presentations of autism are characterized by a common cognitive deficit: an inability to form metarepresentations, arising from some dysfunction in the necessary psychological machinery (the ToMM). As propositional attitudes (i.e., mental states) are metarepresentational, this dysfunction is said to cause ToM deficits in autism (Leslie & Frith, 1987, 1990; Leslie & Thaiss, 1992). Specifically, Leslie (1987) argued that an individual cannot gain an understanding of mental state concepts (necessary for ToM) without the capacity to hold metarepresentations, which itself requires the ability to decouple their propositional content. Therefore, the ToMM account suggests that autistic individuals cannot develop mental state concepts (i.e., informational relations such as "believe" or "desire") due to an impairment in representing propositional content that is decoupled from the real world (Leslie & Thaiss, 1992). Ultimately, the ToMM account suggested that ToM differences in autism arise from a representational deficit and therefore made the testable claim that autism should be characterized by an inability to represent mental states (Baron-Cohen, 1990).

Development Route 1: Explaining the Apparent Nonuniversality of ToM Deficits in Autism

Since the proposal of the ToMM account, much theorizing around the ToM hypothesis of autism has centered on explaining evidence that the vast majority of autistic adults, and many autistic children, are able to pass tasks that require mental state representation (Baron-Cohen, 1989; Baron-Cohen et al., 1985; F. G. Happé, 1995). This evidence has led some to reject the notion that autism is characterized by an inability to represent mental states (Gernsbacher & Yergeau, 2019) but has led others to propose explanations of these data within

the context of the ToMM account (Frith et al., 1991, 1994) and yet led others to develop new theories regarding the nature of the proposed ToM deficit in autism (Senju, 2012). Here, we explain the challenge posed to the ToMM account by evidence that many autistic individuals can pass false belief tasks and other tests of ToM and critically evaluate several attempts to account for this challenge.

The Challenge: Most Autistic Adults, and Many Autistic Children, Pass Tasks Requiring ToM

Initial evidence for the ToMM account was provided by false belief tasks, which are described in detail in Table 2. False belief tests require individuals (usually children) to attribute a belief to a character, despite themselves knowing that the belief is false. When the Sally–Anne task, a classic false belief test, was given to typically developing children, children with Down's syndrome, and autistic children, a high proportion of the former two groups passed (85% and 86% respectively), despite the intellectual impairments associated with Down's syndrome. By contrast, only 20% of the autistic children passed (Baron-Cohen et al., 1985).

The logic behind this form of test suggested that it should not be possible to attribute a false belief to an agent without being able to represent another's mental state as distinct from both reality and one's own mental state (Wimmer & Perner, 1983). Notably, then, it should not be possible to pass the false belief task without the ability to decouple propositions embedded within metarepresentations, raising the question of how, if the ToMM account is correct, 20% of autistic children could pass the task. The importance of the 20% of autistic "passers" was noted at the time, and it was argued by some that the data suggested a developmental delay in the (eventually intact) ToMM in autism (Baron-Cohen, 1989). However, no explanation was offered as to why this delay would impair further development and thus be responsible for social communication difficulties beyond the age at which the impairment is observed.

While 20% of autistic passers is explicable on a single, simple false belief test where chance performance is 50%, in the intervening years, some autistic individuals have been shown to pass (or perform above chance in) all forms of ToM test (see Gernsbacher & Yergeau, 2019, for review). Given that false belief tasks are thought to be highly conservative tests of the ability to represent the mental states of others, these data suggest that, following an initial delay (F. G. Happé, 1995), any difference in ToM in autism is unlikely to be a problem of representation. Therefore, the fact that most autistic adults (and a sizable proportion of autistic children) appear to be able to represent mental states poses a significant problem for the ToMM account, which proposed that a dysfunctional ToMM (and thus an inability to hold metarepresentations) forms the common cognitive explanation of autism across all individuals (Frith et al., 1991).

Proposal 1: ToMM Development Occurs Normally in a Subgroup of Autistic Individuals

Some theorists have responded to data suggesting preserved ToM capacity in autism by suggesting that subgroups of autistic individuals show distinct cognitive differences, each resulting in specific behavioral presentations (Goodman, 1989; F. Happé et al., 2006). For example, a dysfunctional ToMM might only be present in autistic individuals who experience more pronounced social communication difficulties, while symptoms of autism that are less social in nature, such as restrictive and repetitive behaviors, might be better explained by considering executive dysfunction or weak central coherence (F. Happé & Frith, 2006; Lopez et al., 2005). If this is the case, then the lack of universality does not pose a significant problem for the notion that dysfunction in the ToMM explains autism in at least one subgroup of individuals.

However, if it is acknowledged that only some autistic individuals have problems with representing mental states, then the ToMM account ceases to provide a cognitive theory of autism but rather becomes a description of a cognitive impairment seen in some individuals with autism that produces some of their symptoms. Although this limits the scope of what the ToMM account might explain, it is possible that some autistic individuals do have a primary deficit in processing metarepresentations that cause their behavioral symptoms. Such individuals would be expected to fail false belief tasks but perform well on carefully matched control tasks, which do not involve metarepresentation. To our knowledge, no evidence of this exists (i.e., evidence of *individuals* failing a suite of false belief tests while passing carefully matched control tasks). Additionally, the notion of subgroups cannot explain why some individuals perform well in some tasks that require metarepresentation, but not in others, such as in first- and second-order false belief tasks (Baron-Cohen, 1989), or false belief tasks and vignette tasks (F. G. Happé, 1994). As previously discussed, it is not clear how metarepresentational demands could vary across ToM tests, meaning that differences between tests cannot be explained by the ToMM account.

A yet more extreme defense of the original ToMM account—that "passers" are not truly autistic but have been misdiagnosed as autistic—is also possible. It has been argued that the diagnostic criteria used in autism research are both too vague and too broad and thus that the likelihood of Type II error (in which true population group differences are not observed in a given sample) is inflated by the recruitment of autistic participants who may be highly different to each other and relatively similar to the neurotypical sample (Mottron, 2021). While it is possible that this is the case, any diagnostic criterion that explicitly excludes those who can perform ToM tasks is scientifically undesirable, as it would result in the ToM hypothesis becoming untestable.

Specifically, without independent diagnostic criteria, one could explain away any autistic individual shown to have typical ToM as not truly autistic. Indeed, the same criticism may be leveled at the notion that cognitive differences vary between subgroups of autistic individuals. Without an independent method for classifying these subgroups, one could suggest that any individual shown to have typical ToM occupies a distinct subgroup. To consider autism as a single diagnostic category (Regier et al., 2013), one must assume *some* homogeneity, at least in etiology, else distinct subgroups would be better considered as having distinct disorders. Any explanation of autism should, therefore, give some explanation of that homogeneity.

Proposal 2: In the Absence of a Functioning ToMM, Individuals May Pass ToM Tasks Using Simple Behavioral Rules

Other work responded to the ostensible nonuniversality of ToM deficits in autism by arguing that autistic individuals pass tests of false belief understanding through "hacking out" (Frith et al., 1991,

Table 2
A Summary of Existing ToM Tests

Test	Description	Evaluation
False belief tasks	False belief tasks typically take one of two forms: change-of-location/unexpected-transfer tasks or unexpected-contents/deceptive-appearance tasks. In the former (Baron-Cohen et al., 1985; Wimmer & Perner, 1983), an agent is introduced and shown observing or placing an object in a location. The agent then leaves, and the object moves. Participants are asked where the agent will look for the object on their return. In the latter (Gopnik & Astington, 1988; Perner et al., 1989), children are shown a container, which implies (through, e.g., branding or labeling) that it contains a certain object. Participants are then shown that the box contains something unexpected and asked what a naïve friend would think the box contains. In both forms of task, the target of inference is assumed not to have access to information about the true state of the world, and so children are said to pass if they ascribe a false belief to the target.	Thought to be highly conservative tests of the ability to represent mental states, meaning that the ability of many autistic individuals to pass these tasks suggests intact representational capacity in autism. May make significant demands on cognitive abilities other than ToM—particularly executive functions and language ability (Devine & Hughes, 2014; Durrleman & Franck, 2015; Lind & Bowler, 2009; Müller et al., 2005).
Implicit false belief tasks	Implicit false belief tasks are extensions of the false belief tasks described above and usually utilize a change-of-location paradigm. Instead of asking participants explicitly where a character will search for an object, they measure participants' looking behaviors. Anticipatory looking toward the location that the character is assumed to falsely believe the item occupies is said to indicate some implicit understanding of false belief and thus implicit ToM ability (Kovács et al., 2010; Senju et al., 2009; Southgate et al., 2007).	Poor replicability and convergent validity (Kulke & Rakoczy, 2018; Kulke et al., 2018, 2019). At least some effects attributed to automatic belief representation have been shown to occur due to experimental artifacts (Phillips et al., 2015).
Perspective-taking tasks	Some tasks assess "implicit" ToM understanding by examining the automatic processing of others' visual perspectives. The most common example of this type of task is the "dot perspective task" (Samson et al., 2010). In the dot perspective task, participants are asked to verify how many red discs are visible in a scene from either their own or an avatar's perspective. Participants are typically slower and less accurate at verifying their own perspective when it is inconsistent with the avatar's, and this is taken as evidence that they automatically process the avatar's perspective.	Evidence from studies in which the avatar is replaced by a camera, or in which the avatar is unable to see the discs, suggests that these effects occur in the absence of mental state processing and may be better explained as attentional phenomena (Catmur et al., 2016; Conway et al., 2017; Santiesteban et al., 2015). Artificial agents used in these tasks (e.g., avatars) do not truly have mental states.
Strange stories task	The strange stories task (F. G. Happé, 1994) involves participants listening to or reading stories that describe situations designed to elicit attribution of mental or physical states. In the mental state questions, participants are asked to explain a character's behavior, and their responses are scored for accuracy. Other versions of this task involve films depicting the stories (Murray et al., 2017) or modified versions of the original stories for use with children (S. White et al., 2009).	The correctness of responses is determined by whether the raters deem the response "appropriate" as a justification for the story given.
Movie for the assessment of social cognition	The movie for the assessment of social cognition (Dziobek et al., 2006) involves participants watching a video of a social interaction between a group of characters. The video is paused at several time points, and participants are asked multiple-choice questions about the mental states of the characters, as well as control questions that do not ask about mental states but instead assess attention, memory, and inference.	Includes emotion as a type of mental state, in contrast to theoretical definitions of mental states as propositional attitudes. Accuracy is based on a "correct" answer determined by the experimenter's intention and consensus scoring.
Frith–Happé Animations Test	The Frith-Happé Animations Test (Abell et al., 2000; Castelli et al., 2002; S. J. White et al., 2011) involves participants watching a video of two moving shapes. In the original version of this task (Abell et al., 2000; Castelli et al., 2000), participants are asked to describe the video through free verbal response and are scored on the appropriateness of their descriptions and the level of mental state attribution they include. In newer versions (Livingston et al., 2021; S. J. White et al., 2011),	Tests the tendency to attribute mental states to inanimate shapes that cannot hold mental states, not to agents who can. Appropriateness is determined relative to the "scripts" used in the development of the stimuli, that is, what the experimenter intended to depict.

(*table continues*)

Table 2 (*continued*)

Test	Description	Evaluation
	participants are asked to provide a verbal description, which is scored for appropriateness as above, but are also asked to select whether the video is best described as depicting "no interaction," "physical interaction," or "mental interaction."	
Reading the Mind in the Eyes Test	The Reading the Mind in the Eyes Test (Baron-Cohen et al., 1997, 2001) involves participants viewing images of targets' eye regions and selecting, from multiple possible answers, the term that best describes the emotional state or state of mind of the target.	Emotional states or states of mind are distinct from the definition of mental states as propositional attitudes. Accuracy is determined by whether participants select the target word judged to describe the image by the experimenters and the consensus of eight judges.

Note. ToM = theory of mind.

1994). Specifically, it was suggested that, while autistic individuals cannot represent mental states, some may derive a set of behavioral "rules" that outline what another agent will do, or how they themselves should behave, in different situations. This proposal suggests that false belief tasks can be solved using rules relating to the agent's behavior (e.g., "when people do not see a change in the world, they behave as if it has not happened"), or one's own (e.g., "when participating in a psychology study, pick the answer opposite to that which seems most obviously correct"). In such cases, then, the test is thought to be solved without appealing to another's mental state.

Some autistic individuals who pass false belief tasks do show little evidence of ToM use in everyday life (as assessed by observer report based on an expanded version of the Vineland Adaptive Behavior Scales; Frith et al., 1994), and it is possible that these individuals solve ToM tests through "hacking out." However, the same data show that others do make use of ToM day-to-day, and it is hard to maintain that they do so without representing mental states. Specifically, the number of behavioral rules an individual can memorize is likely to be much lower than the number required to substantially influence everyday social communication. Therefore, while the notion of "hacking out" may reduce the number of autistic individuals thought to be truly able to represent mental states, it does not resolve the problem entirely, as some proportion are still thought to be capable of mental state representation. In addition, despite the claim that these individuals still show some ToM impairment, there remains no theorized reason for false belief tasks to be any less demanding of the ToMM than everyday social behaviors (with respect to metarepresentation or representation of propositional attitudes), and so this impairment cannot be explained as a difficulty with metarepresentation.

Proposal 3: The Ability to Explicitly Reason About Mental States May Be Preserved, but Spontaneous, Implicit Mental State Representation Is Impaired in Autism

Another proposal that may account for evidence of preserved ToM ability in autism suggests that while autistic individuals are in fact able to reason about mental states explicitly, they cannot perform unconscious, spontaneous, or "implicit" ToM (Senju, 2012). In this account, it is argued that because the innate ToMM, which processes mental states quickly and spontaneously (an otherwise rarely discussed element of the ToMM account; Leslie, 1994), is missing or dysfunctional in autistic individuals, these individuals can only perform mentalizing tasks through slow, explicit, inferential reasoning. The lack of implicit mentalizing would therefore be expected to mean that autistic individuals have less of a propensity to mentalize, find mentalizing not automatic but effortful, and require attentional focus.

The usefulness of this approach is limited in several ways. First, there are questions surrounding the evidential basis of this account and its theoretical coherence. Empirically, several of the paradigms used to support the notion that implicit ToM is part of typical human cognition have been found to have poor replicability and convergent validity (Kulke & Rakoczy, 2018; Kulke et al., 2018, 2019). Additionally, results thought to be indicative of the presence of implicit ToM may be better explained by other capacities or features of experimental design (Catmur et al., 2016; Cole & Millett, 2019; Conway et al., 2017; Heyes, 2014; Kuhn et al., 2018; Millett et al., 2020; Phillips et al., 2015; Santiesteban et al., 2015). For example, evidence of automatic visual perspective taking (where participants' performance on a task is influenced by what another agent can see, and this is taken to imply automatic processing of others' viewpoints; see Table 2 for more details) can be explained by attentional effects (Catmur et al., 2016), while effects attributed to automatic belief-representation are not sensitive to the agent's belief but associated with the timing of critical agent-related events (Phillips et al., 2015).

At the theoretical level, some of the key approaches to measuring implicit ToM are problematic because they include phenomena such as "seeing" (Apperly et al., 2010; Dumontheil et al., 2010; Keysar et al., 2003; Samson et al., 2010) as mental states. Processing an alternate perspective toward the visual world need not involve the representation of another's mental state—it can be conceived of as a problem of geometry rather than of inferring and representing another's attitude toward the truth of a proposition (Leslie & Frith, 1988). For example, to infer that an avatar sees two red circles on a wall, the observer needs only to determine that there are two red circles in the unobstructed eyeline of the avatar. If tasks involving representation of another's visual perspective were shown to be solved by appealing to the beliefs of the agent, for example, that the avatar believes that there are (only) two red circles on the wall, then these tasks may be considered ToM measures. However, there is evidence to suggest that this is not the case, as the same effects often attributed to implicit ToM are present in control conditions where the agent is replaced by a camera, which cannot hold mental states (Santiesteban et al., 2015), or where the agent is shown to be wearing goggles, which render the visual stimuli invisible to them (Conway et al., 2017).

Second, the notion of disrupted implicit ToM is less able to explain social communication difficulties in autism relative to the original ToMM account. Theories postulating the existence of an implicit ToM argue that such capacities would necessarily be limited and inflexible, denying that the implicit system allows representation of beliefs and instead stating that it merely allows representation of "belieflike" states (Apperly & Butterfill, 2009). Typically, social difficulties ascribed to ToM differences relate to complex social situations, such as the understanding of pragmatic language, sarcasm, or lying (Baron-Cohen et al., 1999; Frith et al., 1994; Reindal et al., 2023; Thiébaut et al., 2016), which likely require the integration of multiple sources of information to reach a ToM judgment, alongside additional cognitive abilities (such as working memory). The use of ToM to reach an accurate inference in these complex situations, at least early in typical social development if not throughout the lifespan, requires the more effortful, but more flexible, explicit capacity (Apperly & Butterfill, 2009). If this capacity, which can facilitate understanding in these complex situations, is thought to be present in autism, then any identified impairment in implicit processing is a poor explanation for autistic symptoms.

Development Route 2: Increasing the Domain Specificity of the Proposed Cognitive Difference

A second key theoretical development concerning the ToM hypothesis is that newer theories tend to suggest that ToM impairments in autism arise from domain-specific social cognitive differences, in contrast to the original ToMM account's suggestion of a general metarepresentational impairment. This increasing emphasis on the social nature of ToM deficits in autism may have originated from data from the false photograph task, which was thought to challenge the notion that ToM difficulties in autism arise from an inability to process *any* metarepresentation (i.e., any representation with embedded propositional content that may not reflect the real world). Here, we outline this challenge, as well as a refocusing of the ToMM account, which emphasized the role of the agent as an essential component of a metarepresentation. We then discuss a broad class of what we term "social processing theories," which also argue for a domain-specific impairment, suggesting that ToM difficulties arise from differences in the processing of social information and thus in the input to the ToM system.

The Challenge: Autistic Children Pass Tasks Claimed to Require Nonmentalistic Metarepresentations

The original ToMM account stated that autistic individuals cannot hold metarepresentations because of an impairment in their ability to decouple the propositions embedded within them (Leslie, 1987; Leslie & Frith, 1987). While the kinds of metarepresentations discussed by Leslie (those involved in pretense and ToM) were said to take the form *agent—informational relation—proposition*, it remained the case that such a deficit could be expected to affect the ability to hold nonmental metarepresentations, not just metarepresentations of mental states (Leslie & Frith, 1990). As such, empirical data suggesting that the metarepresentational deficit in autism is specific to mental state understanding (mentalistic metarepresentation, sometimes called "m-rep") would not support the existence of the domain-general metarepresentational deficit outlined in the original account.

Claims of a metarepresentational deficit in autism that is specific to mentalistic metarepresentations have been made based on data from the false photograph task. The false photograph task, in form, mirrors a false belief task but is claimed to require nonmentalistic metarepresentations for successful performance (Zaitchik, 1990). In the false photograph task, a photograph is taken of a group of objects. The objects are then moved away from their original location, and the participant is asked "In the picture, where is the object?" Just as in a false belief task, where children are said to "pass" if they can recognize that a character's beliefs do not match reality, children are said to "pass" the false photograph test if they recognize that the location of the objects in the photograph does not match (present) reality.

Leslie and Thaiss (1992) presented evidence that autistic children perform well on false photograph tasks despite their failure on the false belief task, while typically developing 4-year-olds perform poorly on the false photograph task but well on the false belief task. They argued that these data show that the false photograph task is not easier than the false belief task as the typically developing children appeared to find it more challenging. As such, if one takes the processing of a false photograph to require nonmentalistic metarepresentation, then this evidence refutes the suggestion by the early ToMM account that ToM difficulties in autism arise from a domain-general metarepresentational deficit.

Whether data from the false photograph task are, in fact, problematic for the original account is a topic of some debate (Frith et al., 1991; Iao & Leekam, 2014; Leslie & Frith, 1990). Photograph understanding could be conceived as requiring representation of an opaque proposition as "the photo shows the marble is in the basket" does not imply that, in the real world, the marble is currently in the basket. Therefore, the false photograph task could require decoupling and be conceived of as a nonmentalistic metarepresentational task. However, the fact that a photograph does entail facts about how the world *was*, even if not necessarily about how it *currently is*, might mean that photographs are not truly metarepresentational (Perner & Leekam, 2008). Under this view, the false photograph task does not require metarepresentation, and the photograph may be more analogous to one's own memory than to another's belief. Thus, autistic children who pass this task may not necessarily be processing metarepresentations to do so, and, as such, false photograph data may not challenge the original ToMM account.

In fact, other works suggest that autistic children may indeed show difficulties in nonmentalistic metarepresentation. For example, work by Iao and Leekam (2014) demonstrates that autistic children are impaired at the "false sign task," another task analogous to false belief tasks for which it can be claimed with a great deal of certainty that nonmentalistic metarepresentation is required for successful performance. In this task, a physical sign misrepresents the *current* location of an object (by pointing at an incorrect location) and thus cannot be characterized as an outdated, but once veridical, representation in the way that a photograph might be (Bowler et al., 2005). Therefore, the metarepresentation in this case certainly does require decoupling. The fact that autistic children are more likely to be impaired at the false sign task than typically developing children—as they are at false belief tasks—is consistent with the original account of a decoupling impairment leading to a general metarepresentational deficit in autism.

Ultimately, while false photograph data were originally thought to challenge the notion of a general metarepresentational deficit in

autism, it appears that the task may not actually require metarepresentation and thus that the data do not challenge the original ToMM account. However, the previously described evidence that a large proportion of autistic individuals do develop the ability to represent mental states (if passing false belief tasks show the ability to represent mental states) demonstrates that any general metarepresentational deficit does not often persist into adulthood and that the ToMM account thus provides a poor explanation of autism beyond childhood. The relevance of the false photograph challenge, then, lies in the fact that, regardless of its validity, it prompted revisions to the original ToMM account, which have continued to influence theorizing around ToM in autism.

Proposal 1: The ToMM Is Responsible Only for the Understanding of Mental States and Is Damaged in Autism

In response to the perceived challenge of the false photograph data, Leslie and Thaiss (1992) suggested that autistic individuals have a specific problem with mentalistic metarepresentations. They argued that there can be no informational relation in the false photograph task because the photograph (or the camera) is not an agent with a mind, so cannot *believe* the marble to be in the basket. Additionally, Leslie (1994) further argued that the ToMM is specialized for the processing of intentional (mental) states. We consider these suggestions to constitute a revision to the original ToMM account because the claim that deficits in autistic ToM arise from *some* dysfunction in a module responsible *only* for mental state understanding is inconsistent with the original formal theory, which suggested that the difficulty is one of processing decoupled propositional content.

The original conceptualization of the ToMM account claimed to explain ToM deficits in autism by suggesting that ToM required the processing of a form of representation that required specific representational abilities (i.e., the ability to process decoupled propositional content). As a consequence of the lack of the ability to process these representations, mental states (among other metarepresentations) could not be represented. By contrast, the revision to the ToMM account that specifies that autistic individuals have a selective deficit in representing mental states means the theory can no longer *explain* ToM deficits in autism; it merely redescribes ToM deficits ("autistic individuals show a deficit in ToM because they cannot represent mental states"). This revised version of the ToMM account therefore cannot offer a satisfactory theoretical explanation of ToM in autism.

Proposal 2: Metarepresentational Mechanisms Are Intact, but Social Processing Deficits Cause Impoverished Input

Another class of theory, which we term "social processing theories," also suggests that ToM ability is supported by domain-specific processes that are thought to be impaired in autism. Such theories might therefore follow from the revised version of the ToMM account outlined above, providing possible explanations of the proposed domain-specific impairment. These theories suggest, in contrast to the ToMM account, that ToM deficits arise from difficulties attending to and processing the social information required for ToM inference (Stone & Gerrans, 2006), or from difficulties developing social skills thought to be necessary precursors to the emergence of ToM ability (Garfield et al., 2001).

While there is a substantial (but contested, see Bottema-Beutel et al., 2019, for a relevant meta-analysis) literature pointing to differences in social processing in autism, such as in social orienting and joint attention (Burnside et al., 2017; Dawson et al., 2004; Senju & Johnson, 2009a, 2009b; Shah et al., 2013), there is little work outlining how exactly these differences in social processing relate to ToM ability. Evidence that certain social abilities are correlated (cross-sectionally or longitudinally) with ToM (Burnside et al., 2017; Charman et al., 2000; Sodian & Kristen-Antonow, 2015) does not in itself explain how these social abilities are thought to be necessary for the inference of mental states. There is clearly a need for a psychological model explaining precisely how observed atypical processing of social information in autism impacts mental state inference ability.

Ultimately, then, while social processing theories can account for evidence that autistic individuals appear able to hold metarepresentations, they do not provide specific, testable hypotheses about the mechanisms underlying ToM and how they might differ in autism. They do, however, provide a promising starting point for the development of such hypotheses, as one might begin to conceptualize a mechanistic theory of ToM by considering the possible roles of social processes identified as operating differently in autism.

Developments in Theoretical Proposals Have Not Provided a Satisfactory Psychological Model of ToM in Autism

In this section, we have traced the development of the ToM hypothesis of autism from its origins in the ToMM account to the present day, examining developments that either sought to account for evidence of ToM ability in (some) autistic individuals or proposed that ToM deficits arise from domain-specific social cognitive impairments rather than domain-general metarepresentational difficulties. As we argued in Defining the Conceptual Space of ToM in Autism section, we believe that a satisfactory psychological model of ToM in autism should specify mechanisms underlying mental state inference, the information transformed through those processes, and how those processes might differ between autistic and neurotypical individuals. Each theoretical account that we have described faces its own challenges, and none, in our opinion, meets these criteria. A summary of the proposals evaluated in this section and what we consider to be their problems is given in Table 3.

Existing ToM Tests Cannot Conclusively Support the ToM Hypothesis

Thus far, we have established some conceptual difficulties that arise when theorizing about ToM and critically evaluated existing proposals, concluding that none provides a satisfactory account of ToM differences in autism. To identify challenges that may be faced when testing any new theoretical account that may emerge, we turn now to evaluating existing tests of ToM ability. In general, tests of ToM ability may be considered to test mental state representation, the accuracy of mental state inferences, and/or the propensity to make mental state attributions. In this section, however, we will argue that existing tests have significant limitations that may prevent them from conclusively identifying differences in any of these aspects of ToM between autistic and neurotypical individuals.

Table 3
A Summary of Proposals Discussed in the Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism Section and Their Problems

Proposal	Summary	Problem
ToMM account	An innate neural module is responsible for processing propositions that need not be reflective of reality (as "metarepresentations"), including propositional attitudes (mental states). Autism is characterized by dysfunction in this module and an inability to hold these types of representation.	Possible circularity. Empirical data do not support the idea that autistic individuals cannot hold these forms of representations.
Subgroups	The ToMM account holds for some subgroups of autistic individuals, but not for others. Different subgroups may show different cognitive profiles.	Reduces the extent to which the ToMM account is able to explain autism as a single condition. Cannot account for autistic individuals who pass simple, but not "advanced" tests. Renders ToMM account untestable
"Hacking out"	Some autistic individuals, despite the lack of the necessary innate module for ToM (as in the ToMM account), are able to pass laboratory tests of ToM through logical reasoning that does not require the representation of mental states.	Cannot account for evidence of ToM use in everyday life in some autistic individuals.
Implicit ToM	Autistic individuals are able to perform slow, explicit mental state reasoning, but not fast, automatic, implicit ToM.	Does not specify how implicit ToM differences occur or how differences affect social communication abilities. Questions surrounding the validity of the evidential basis
ToMM (revised)	The innate neural module thought to be dysfunctional in autism is responsible only for processing propositional attitudes (i.e., propositions that may not be reflective of reality *and* that an agent has an attitude toward). Therefore, autistic individuals can hold some forms of metarepresentation but cannot represent mental states.	No longer proposes an explanation for processing differences underlying ToM impairments but simply redescribes them. Empirical data do not support the idea that autistic individuals cannot hold these forms of representations.
Social processing	The neural or psychological system required for ToM is intact in autism, but ToM impairments arise from differences in the social information that forms the input to the system. This is due to differences in social processing or social behavior in autism (e.g., in joint attention or social orienting).	Do not provide explanations of how differences in some social processes lead to differences in ToM judgments. Questions surrounding validity of evidential basis.

Note. ToMM = theory of mind module; ToM = theory of mind.

A summary of existing tests and their limitations can be found in Table 2.

We begin by evaluating the extent to which advanced tests can be thought to be testing ToM and only ToM. We then consider, in turn, measures of each aspect of ToM (representation, propensity, and inference accuracy), exploring specific problems facing each type of test. Ultimately, we conclude by identifying requirements of any new ToM measure that seeks to test differences in ToM ability between autistic and neurotypical individuals.

Many Tests Do Not Solely Test ToM Ability

As explained in the Defining the Conceptual Space of ToM in Autism section (and summarized in Table 1), ToM may be defined as the ability to infer mental states, with mental states themselves defined as propositional attitudes. A propositional attitude is an agent's attitude to the truth of a particular proposition, and these attitudes are typically beliefs, desires, or intentions. Other forms of mental event, such as emotion, visual perspective, or state of mind, are distinct from propositional attitudes.

In addition to tests of visual perspective taking which, as discussed in the Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism section, do not require consideration of mental states and often appear to be solved without processing the agents' visual perspective (Conway et al., 2017; Leslie & Frith, 1988; Santiesteban et al., 2015), other tests also stray from the use of propositional attitudes. For example, the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001) assesses the identification of facial expressions of emotion or states of mind (nonaffective states that the mind may occupy, such as "reflective" or "thoughtful"), which do not hold propositional content. As such, without an expanded definition of what constitutes a mental state, this test does not appear to test ToM ability. Similar criticisms can be made of the Movie for the Assessment of Social Cognition (MASC; Dziobek et al., 2006), which includes emotion as a "mental state modality" alongside propositional attitudes, meaning that the MASC may conflate emotion identification with the inference of mental states (see Shah et al., 2017, for evidence of differential associations between the emotional and nonemotional items on the MASC).

Importantly, the inclusion of emotion inferences in tests of ToM may disadvantage autistic participants due to the frequent co-occurrence of autism and alexithymia, a condition known to impair emotion processing (Bird & Cook, 2013; Hill et al., 2004; Sifneos, 1973). Concerningly, then, group differences between autistic and nonautistic participants in tests that include emotion judgments may be better explained by group differences in alexithymia rather than autism itself. Moreover, like group differences in visual perspective tasks, which may be explained by lower level, domain-general abilities (Catmur et al., 2016; Cole & Millett, 2019; Conway et al., 2017; Millett et al., 2020; Phillips et al., 2015; Santiesteban et al., 2015), group differences in emotion judgments might not reflect a ToM difference even if autism specific (Oakley et al., 2016).

Tests of Representation: Differences in BOLD Signal When Representing Mental States Cannot Be Properly Interpreted

Although, as we argued in the Defining the Conceptual Space of ToM in Autism section, it is unclear how there may be degrees of the ability to represent, it is frequently claimed that evidence of a reduced ability to represent mental states in autism can be found in fMRI studies. Such studies have shown reduced BOLD signal in a network related to mental state representation in autistic individuals (Arioli et al., 2021; Kana et al., 2015; Nijhof et al., 2018).

As we discussed in the Defining the Conceptual Space of ToM in Autism section, any BOLD signal unique to a condition that requires mental state representation, relative to a carefully matched control condition, may be reflective of the neural processes underlying mental state representation. As such, differences in this BOLD response could be indicative of differences in mental state representation across groups. However, some of the better powered studies investigating neural differences between autistic and neurotypical individuals during ToM tasks have failed to find any differences (Dufour et al., 2013; Moessnang et al., 2020). While the empirical evidence is therefore mixed, even if there is reduced neural activity in a network activated during ToM tasks in autism, it is unclear what can be concluded about mental state processing on this basis. One could argue that reduced neural activity indicates a weaker representation of mental states or that reduced neural activity indicates that the representation of mental states is occurring more efficiently. Thus, without accompanying behavioral data, it is almost impossible to make conclusions about psychological representations from these studies. Several studies that do show differences in BOLD signal between autistic and neurotypical groups fail to find significant behavioral differences in their ToM task (Kana et al., 2015; Nijhof et al., 2018), further confusing the possible interpretation of these results.

Furthermore, it has been convincingly argued that BOLD should not be used to make inferences about differences in neural activity between autistic and neurotypical individuals (Reynell & Harris, 2013) due to differences in neurovascular coupling between these groups. As far as we are aware, there have been no fMRI studies, including our own (Silani et al., 2008; Spengler et al., 2010), that have accounted for such differences, and it is unclear whether it is possible to do so. Having identified differences in neurovascular coupling as a potential confound, Reynell and Harris highlighted that these differences can be pathway specific and therefore task specific. As such, if one assumes that there are specific neural mechanisms for mental state inference, which are distinct from those used in inference more generally, then even a well-matched control task cannot rule out this potential confound.

Tests of Inference Accuracy: Existing Tests Do Not Have a True Correct Answer

As explained in the Defining the Conceptual Space of ToM in Autism section, if ToM is a process of inference (i.e., of combining available information to reach a judgment about a target individual's mental state), then one way in which individuals or groups might differ in their ToM ability is in the accuracy of these inferences. Several tests of ToM ability purport to be testing ToM accuracy. In such tests, participants are asked to infer mental states, either with a free response or by selecting from multiple response options (Abell et al., 2000; Dziobek et al., 2006; F. G. Happé, 1994). These mental state inferences are then assessed for their accuracy against a predefined "correct" answer. Often, however, these inferences are to be made about the minds of targets in artificial stimuli, for example, written vignettes or videos involving actors. In such cases, the target does not truly hold the mental state being assessed. An actor likely does not hold the mental states attributed to their character by the experimenter but merely performs behaviors that they believe the character would display when experiencing those mental states. Similarly, in vignettes written by the experimenter, the "mental state" is what the experimenter intended to depict, not the true experience of any real individual.

This lack of a true mental state causes problems in assessing the accuracy of participants' inferences given that, at a basic level, it is impossible to assess the accuracy of an inference unless there is a true answer. In these tasks, the "correct" answer is instead defined as that which the experimenter intended to depict (regardless of how they did so or how successfully) or the most frequent answer provided by a large group of neurotypical raters (Dziobek et al., 2006). Problematically, if an autistic participant responds differently to experimental task demands, then they may give a different response to the neurotypical consensus group even if they have no ToM impairment. For example, neurotypical individuals may interpret the instruction to infer mental states from the stimuli as a request to infer what the experimenter intended to depict and make use of dramatic conventions in their judgment. By contrast, autistic individuals may take the question literally, and struggle to ascribe a mental state to a character, in the knowledge that the character is a product of a script and the actor's performance and thus does not truly hold a mental state.

Logically, if one is to argue that differences in ToM underlie difficulties in social communication and understanding social pragmatics in autism, then one should not test for ToM impairments using an experimental situation where different (pragmatic) interpretations of the task instructions could lead to differences in measured ToM accuracy. Given that difficulties with social pragmatics are often observed in autism, it is likely that autistic individuals will interpret task instructions differently from neurotypical individuals and, in turn, make different inferences to neurotypical individuals. If a neurotypical consensus group is used to define the correct answer, then "inaccurate" inferences made by autistic individuals may be reflective only of their differing interpretation of task requirements rather than impairments in mental state inference itself. As such, tests using this approach are biased toward neurotypical cognition and thus cannot truly test group differences in ToM ability.

Tests of Propensity: Existing Measures Do Not Test (Only) the Attribution of Mental States to Social Agents

In addition to ToM accuracy, the propensity to spontaneously attribute mental states to other agents, or to use those mental states to explain or predict behavior, may be an important component of ToM ability, and reduced propensity could explain some of the social communicative symptoms of autism. If one is not inclined to represent another's mental states and the relationship between them and the content of the individual's speech, one may, for example, be

poor at detecting sarcasm and may instead tend to take the content of utterances literally.

The propensity to consider mental states is generally tested by examining whether mental state terms are used spontaneously when explaining or describing a situation (Abell et al., 2000; F. G. Happé, 1994). The most popular measure of propensity, the Frith–Happé Animations Test (Abell et al., 2000; Castelli et al., 2002; Kana et al., 2015), involves describing the behavior of animated triangles. In the original Abell et al. (2000) scoring system, verbal responses were categorized as "action" (if the participant did not offer any explicit reference to the interaction between the triangles), "interaction" (if they referred to interaction but did not use any psychological terms), or "mentalizing" (if the participant used psychological terms). It is this categorization that measures participants' propensity to make mental state attributions. In more recent versions of this task (Livingston et al., 2021; S. J. White et al., 2011), participants are asked to select whether the animation depicts "no interaction," "physical interaction," or "mental interaction." While this multiple-choice categorization is often considered to measure accuracy (as the experimenters predefine the correct interaction type), it could be argued to be a measure of the propensity to view an interaction as mentalistic (and thus to attribute mental states), especially when prompted that it might be appropriate to do so.

Problematically, the Animations Test examines the propensity to attribute mental states to objects that do not have minds. As such, it is unclear whether this test, which effectively assesses the tendency to anthropomorphize shapes, truly measures the tendency to ascribe mental states to social agents. This may be especially problematic when testing autistic individuals, who have been suggested to show difficulties with generative imagination (Crespi et al., 2016; Low et al., 2009; Ten Eycke & Müller, 2015). It may be the case that a preference for describing things in literal terms impedes the attribution of mental states to objects that cannot hold them, but not to agents who can.

The other primary source of evidence suggesting a reduced propensity to make spontaneous mental state attributions in autism comes from the implicit ToM literature, described in the Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism section. It has been found that autistic individuals are less likely to show anticipatory-looking behavior toward the location where an agent will search in a false belief paradigm (Senju, 2012). As previously mentioned, paradigms like these have been shown to have poor replicability and convergent validity (Kulke & Rakoczy, 2018; Kulke et al., 2018, 2019). Moreover, without a control condition in which the only differing feature is the presence or absence of relevant mental states, group differences in anticipator-looking behavior may be explained by factors other than ToM propensity. Specifically, there is evidence to suggest that autistic individuals show atypical gaze patterns toward social stimuli even in the absence of mental state information (Boraston & Blakemore, 2007) and that both social and nonsocial attention differ between autistic and neurotypical individuals (Bedford et al., 2014). In addition, it has been found that when autistic individuals maintain attention to relevant stimuli, no group differences in goal prediction are observed (Marsh et al., 2015). As such, anticipatory-looking paradigms are limited by the lack of control conditions, which may rule out these potential confounds and may therefore be poor measures of ToM propensity.

New ToM Tests Are Needed

In this section, we have outlined the limitations of existing paradigms for testing ToM and thus for testing ToM differences between autistic and neurotypical individuals. We have noted that ToM testing suffers from the lack of shared understanding of what ToM entails outlined in the Defining the Conceptual Space of ToM in Autism section. Because of discrepancies between how ToM is conceptualized in theoretical terms (as the ability to infer propositional attitudes) and how it is tested (as the ability to infer any of several distinct types of mental event), differences between autistic and neurotypical individuals in performance on several ToM tests may in fact be representative of group differences in abilities other than ToM itself.

We have additionally explained how many ToM tests disadvantage autistic individuals in a manner unrelated to any ToM impairments that may be present. Examples of this include the failure to account for differences in neurovascular coupling in functional MRI studies, the definition of accuracy relative to neurotypical consensus- or experimenter-defined norms, or the requirement, in tests of propensity, to attribute mental states to entities that do not have minds. Failure to account for these possible issues, many of which relate to suspected differences between autistic and neurotypical cognition, limits the extent to which these paradigms can conclusively support the hypothesis that ToM differences are characteristic of autism. As such, to properly test any new theoretical explanation of ToM in autism, new tests will be required.

Reconceptualizing the ToM Hypothesis of Autism: A Worked Example

Throughout this article, we have outlined several issues facing the ToM hypothesis of autism at both the conceptual and empirical levels. We have argued that existing theoretical proposals are insufficient explanations of ToM in autism and that existing ToM tests cannot conclusively support the ToM hypothesis of autism. Here, we discuss our own attempts to reconceptualize the ToM hypothesis of autism. We discuss our model of typical ToM and how it might be used to explain the proposed ToM difficulties in autism. We also describe a measure of mental state inference accuracy, which we believe resolves some of the issues discussed in the Existing ToM Tests Cannot Conclusively Support the ToM Hypothesis section. It should be noted that this work represents only one way in which these challenges may be tackled and does not yet offer a complete explanation of ToM in autism.

The Mind-Space Framework

As explored in the Defining the Conceptual Space of ToM in Autism section, given the evidence of a preserved capacity for the representation of mental states in autism (Dziobek et al., 2006; Frith et al., 1994; Gernsbacher & Yergeau, 2019; F. G. Happé, 1994, 1995; S. White et al., 2009, 2011), we see two candidate aspects of ToM ability that might differ between autistic and neurotypical individuals: the propensity to make mental state inferences and the accuracy of those inferences (Pisani et al., 2021). A new theoretical model of ToM in autism, then, might consider when, how, and why mental state inferences may be less accurate or less often spontaneously produced in autistic individuals. One way of doing this is through

developing a model of typical ToM that identifies potential sources of group differences in mental state inference. The Mind-space framework, a cognitive account of mental state inference (Conway et al., 2019), identifies several mechanisms underlying typical ToM that could operate differently in autism and thus may provide a useful starting point for explaining ToM differences in autism.

The Mind-space framework is founded on the notion that different minds will give rise to different mental states in the same situation. For example, an extraverted individual at a party might intend to speak to as many people as possible—an intention that an introverted person is unlikely to share. The Mind-space framework thus suggests that information about a target's traits (such as their personality traits or cognitive abilities) is used when inferring their mental states. This information is represented as a location in "mind space": a multidimensional space in which each dimension represents a separate trait, akin to how face space allows qualities of faces to be represented (Valentine et al., 2016). Crucially, trait dimensions within this space need not be orthogonal. For example, an individual might have learned that a person's level of suspiciousness is predictive of that person's level of aggressiveness. If the structure of the individual's mind space accurately captures the covariation between these trait dimensions, then this should lead to an improved ability to locate individuals accurately in mind space. Having located their target in mind space, a mentalizer can then use a set of experience-dependent learned relationships to probabilistically infer the likely mental state experienced in this situation by an individual who occupies this mind-space location.

Several possible sources of individual and group differences in mental state inference thus emerge from consideration of the Mind-space framework. These include, among others, (a) the accuracy of the structure of mind space (i.e., the covariance between trait dimensions) relative to population trait covariance, (b) the ability to locate target minds accurately within mind space (which might itself be affected by the ability to attend to and process others' behavior), (c) the ability to make accurate mental state inferences by combining situational information with the target's location in mind space, and (d) the propensity to locate the target in mind space prior to mental state inference. Existing evidence supports some of these ideas, suggesting that individual differences in mental state inference are related to the accuracy of trait dimension covariance in the individual's mind space and the accuracy with which mentalizers locate targets in mind space (Conway et al., 2020; Long et al., 2022), although performance on these tasks in autism has not yet been tested.

We believe that the Mind-space framework resolves many of the issues outlined throughout this article. In this model, mental states are defined purely as propositional attitudes, although aspects of the framework might reasonably be applied to other mental events—one's traits might, for example, also influence one's emotional experiences. By clearly specifying the target of explanation as the accurate inference of propositional attitudes, the Mind-space framework resolves the conceptual issues described in the Defining the Conceptual Space of ToM in Autism section. Additionally, in suggesting specific mechanisms underlying mental state inference, the framework offers potential explanations for evidence that autistic individuals may show differences in ToM despite being able to represent mental states. As such, this framework can account for evidence of preserved ToM capacity without classifying autistic individuals into distinct subgroups (rendering the ToM hypothesis untestable) or relying on the uncertain empirical and conceptual grounding regarding the existence of implicit ToM. Similarly, this mechanistic understanding of ToM might facilitate an understanding of how cognitive differences (domain specific or domain general) result in differences in mental state inference. Differences in social orienting in autism (Burnside et al., 2017; Dawson et al., 2004; Senju & Johnson, 2009a, 2009b), for example, may result in less information being available for locating a target in mind space, causing inaccurate representation of the target's traits and, in turn, inaccurate inference of their mental states. In these ways, the Mind-space framework resolves the issues faced by existing accounts, which we identified in the Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism section.

An additional benefit of this model is that it provides a possible explanation for a set of ideas collectively described under the umbrella of the "double-empathy problem." In describing this problem, Milton (2012) claimed that social content is produced in tandem by interacting agents and that, therefore, individuals with different dispositions (such as autistic and nonautistic individuals) may have different understandings of social norms and behavioral expressions. If this is the case, tests based on neurotypical mental states may disadvantage autistic individuals, falsely suggesting a cognitive deficit. The apparent deficit may derive from a properly functioning ToM system that is unable to compensate for differences in social understanding and expression between differently disposed individuals. This notion is supported by evidence that neurotypical individuals show poor performance when making ToM judgments about stimuli produced by autistic individuals (Edey et al., 2016).

Unlike the existing accounts described in the Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism section, the mind space account can explain ToM differences in autism without assuming the presence of a cognitive impairment and is thus not inconsistent with evidence of double-empathy effects. Specifically, the Mind-space framework suggests that, due to the vast amount of experience gained from privileged access to one's own mind, it should be more difficult to locate a dissimilar other in mind space and to infer their mental state (Conway et al., 2019). This idea is evidenced by data showing a similarity effect in which neurotypical individuals are less accurate when inferring the traits of dissimilar (but still neurotypical) others (Conway et al., 2020). It could be the case, then, that autistic individuals do not have an "impairment" in the cognitive system underlying ToM but instead, given the same experience-dependent machinery, are disadvantaged in a world composed primarily of neurotypical individuals, who may think and behave more similarly to each other than to autistic individuals.

The Mind-space framework, then, makes clear predictions about possible ways in which ToM inference might differ between autistic and neurotypical minds. Empirical work is required to determine whether any of the identified processes do differ in autism, and there is scope to further specify this theory, for example, by exploring the roles of different aspects of the social input (e.g., facial affect, verbal content, body language) or different cognitive abilities, for example, metacognition. Ultimately, however, the Mind-space framework offers a starting point for a reconceptualization of the ToM hypothesis of autism and appears to resolve the issues facing existing theoretical accounts.

The Interview Task

Of course, the development of new psychological models can only aid in our understanding of autism insofar as we can test them empirically. We have argued, in the Existing ToM Tests Cannot Conclusively Support the ToM Hypothesis section, that existing ToM tests have limitations that prevent them from conclusively supporting the notion that ToM differences do occur in autism. Here, we discuss a new test of ToM accuracy, the Interview Task (Long et al., 2022), which we believe to be free of these limitations. This test can be used to measure the accuracy of mental state inferences and can thus identify differences in ToM accuracy between autistic and neurotypical participants.

In the Existing ToM Tests Cannot Conclusively Support the ToM Hypothesis section, we highlighted issues around the determination of the correct answer in tests of ToM accuracy. As we explained, the "correct" answer in such tasks is usually defined either as the mental state the experimenter intended to be communicated or as the response typically provided by a neurotypical consensus group. In these cases, both what the participant is asked to do (report the mental state of a character who does not truly exist) and how they are scored (by the typicality of their reports) mean that "incorrect" answers may not truly indicate deficits in accurate mental state inference. It is necessary, then, to reconsider the scoring of ToM accuracy. Given that an accurate ToM inference is one that matches the beliefs, desires, and/or intentions of the target individual, then the correct answer for such an inference must be the true mental state of the target individual. This cannot be achieved using artificial stimuli that make use of actors or imaginary characters, as the characters in these stimuli do not have minds and, therefore, have no true mental state.

The Interview Task (Long et al., 2022) was designed to meet this requirement. In the Interview Task, participants watch videos of a practice job interview and report their inferences about the traits and mental states of the interviewer and the candidate. Importantly, the interviewers and candidates were not actors but were instead recruited as study participants. Interviewers and candidates engaged in the unscripted practice job interviews, which formed the Interview Task stimuli, and reported their mental states along the same dimensions as participants later used to report their inferences. For example, the candidate was asked "Did the interviewer seem attentive?" and task participants were asked "Does the candidate think that the interviewer seemed attentive?" As such, participant inferences about the mental states of the interviewers and candidates can be directly compared with ground-truth information about the true mental states, and an error score can be obtained. The Interview Task can also measure trait inference accuracy against the true traits of the targets, obtained using validated personality questionnaires and intelligence tests, which is important for testing the predictions of the Mind-space framework.

Participants in the Interview Task are scored in a manner consistent with the instructions given (i.e., inferences about an individual's mental states are scored relative to those mental states, not a consensus group's typical belief about those mental states), and the task truly evaluates the accuracy of the inference process rather than its typicality. Additionally, the Interview Task considers mental states solely as propositional attitudes and so operationalizes ToM in a manner consistent with the proposed theory. The Interview Task therefore resolves issues facing existing measures of ToM accuracy. However, this is only one example of how this can be achieved. For example, the Interview Task makes use of highly naturalistic stimuli—it is likely that the videos contain a multitude of potential cues to the traits and mental states of the interviewers and the candidates, such as the prosody of their voice, the verbal content of their utterances, or their facial expressions. Other researchers might wish to examine ToM accuracy based on more constrained stimuli to isolate the ability to interpret one set of cues, for example. Crucially, however, researchers can only be certain that they are truly measuring ToM accuracy if participants' inferences about a target's mental state are scored against true values provided by the target themselves, whose behavior is presented in the stimuli.

Conclusion

The ToM hypothesis, which suggests that social communication difficulties observed in autism can be explained by differences in mental state inference, holds remarkable potential as a unifying cognitive explanation for social difficulties across heterogeneous presentations of autism. However, the value that this hypothesis offers to our understanding of autistic cognition is limited by the extent to which proposed ToM differences can be described and explained. Indeed, reliance on the generalized notion that autistic individuals "lack ToM," despite its inconsistency with existing evidence (Gernsbacher & Yergeau, 2019), might explain why interventions based on the ToM hypothesis have shown limited success to date (Begeer et al., 2011, 2015; Fletcher-Watson et al., 2014; Lecheler et al., 2021). To fully understand ToM in autism, one requires a clear definition of the target of explanation (i.e., of what ToM is), a theory that specifies the psychological processes underlying ToM and how they may differ in autism, and appropriate measures to test the predictions of the theory. In this article, we have argued that the existing literature fails to satisfy each of these requirements.

In the Existing Theoretical Proposals Are Insufficient Explanations of ToM in Autism section, we traced the development of psychological theory, demonstrating that existing models provide insufficient explanations of any ToM impairment in autism. Specifically, we claim that theories, which posit that autism is characterized by an inability to represent mental states, are inconsistent with empirical evidence. Other models, such as those we described as "social processing theories," require further specification of how the processes they identify are involved in mental state inference and thus how impairments in these processes result in ToM impairments in autism. In our discussion of the mind-space theory, in the Reconceptualizing the ToM Hypothesis of Autism: A Worked Example section, we showed that a mechanistic cognitive model of typical mental state inference might be extended to identify potential sources of ToM differences in autism.

We elaborated, in the Existing ToM Tests Cannot Conclusively Support the ToM Hypothesis section, on issues with existing measures that prevent proper testing of psychological accounts of ToM. We highlighted conceptual inconsistencies between theory and evidence, which mean that tests often measure abilities that do not constitute ToM. We then examined tests of different aspects of ToM ability in turn, arguing that functional MRI-based tests of mental state representation cannot be properly interpreted, that tests of accuracy require a truly correct answer (which is absent from existing measures), and that tests of propensity should only require attribution of mental states to agents with minds (rather than objects

that cannot hold mental states). The test we presented in the Reconceptualizing the ToM Hypothesis of Autism: A Worked Example section, the Interview Task (Long et al., 2022), determines participant inferences against ground-truth information about real mental states and thus resolves the problems facing existing tests of ToM accuracy. However, further work developing a new measure of ToM propensity may be required to facilitate a full understanding of ToM in autism.

Ultimately, by expanding upon the issues that we have identified, and describing how our work attempts to solve them, we hope to encourage others to consider these issues when studying ToM in autism. Such work may not only facilitate greater academic understanding of autistic cognition but may also support the development of interventions that reduce the difficulty and distress sometimes associated with social interaction in autism.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Andrés-Roqueta, C., & Katsos, N. (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. *Frontiers in Psychology*, *8*, Article 996. https://doi.org/10.3389/fpsyg.2017.00996

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Apperly, I. A., Carroll, D. J., Samson, D., Humphreys, G. W., Qureshi, A., & Moffitt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *63*(6), 1201–1217. https://doi.org/10.1080/17470210903281582

Arioli, M., Cattaneo, Z., Ricciardi, E., & Canessa, N. (2021). Overlapping and specific neural correlates for empathizing, affective mentalizing, and cognitive mentalizing: A coordinate-based meta-analytic study. *Human Brain Mapping*, *42*(14), 4777–4804. https://doi.org/10.1002/hbm.25570

Arslan, B., Hohenberger, A., & Verbrugge, R. (2017). Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLOS ONE*, *12*(1), Article e0169510. https://doi.org/10.1371/journal.pone.0169510

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *30*(2), 285–297. https://doi.org/10.1111/j.1469-7610.1989.tb00241.x

Baron-Cohen, S. (1990). Autism: A specific cognitive disorder of "mind-blindness". *International Review of Psychiatry*, *2*(1), 81–90. https://doi.org/10.3109/09540269009028274

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *38*(7), 813–822. https://doi.org/10.1111/j.1469-7610.1997.tb01599.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. https://doi.org/10.1023/A:1023035012436

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Bedford, R., Pickles, A., Gliga, T., Elsabbagh, M., Charman, T., Johnson, M. H., & the BASIS Team. (2014). Additive effects of social and non-social attention during infancy relate to later autism spectrum disorder. *Developmental Science*, *17*(4), 612–620. https://doi.org/10.1111/desc.12139

Begeer, S., Gevers, C., Clifford, P., Verhoeve, M., Kat, K., Hoddenbach, E., & Boer, F. (2011). Theory of Mind training in children with autism: A randomized controlled trial. *Journal of Autism and Developmental Disorders*, *41*(8), 997–1006. https://doi.org/10.1007/s10803-010-1121-9

Begeer, S., Howlin, P., Hoddenbach, E., Clauser, C., Lindauer, R., Clifford, P., Gevers, C., Boer, F., & Koot, H. M. (2015). Effects and moderators of a short theory of mind intervention for children with autism spectrum disorder: A randomized controlled trial. *Autism Research*, *8*(6), 738–748. https://doi.org/10.1002/aur.1489

Bird, G., & Cook, R. (2013). Mixed emotions: The contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry*, *3*(7), Article e285. https://doi.org/10.1038/tp.2013.61

Boraston, Z., & Blakemore, S.-J. (2007). The application of eye-tracking technology in the study of autism. *The Journal of Physiology*, *581*(3), 893–898. https://doi.org/10.1113/jphysiol.2007.133587

Bosco, F. M., Tirassa, M., & Gabbatore, I. (2018). Why pragmatics and theory of mind do not (completely) overlap. *Frontiers in Psychology*, *9*, Article 1453. https://doi.org/10.3389/fpsyg.2018.01453

Bottema-Beutel, K., Kim, S. Y., & Crowley, S. (2019, February). A systematic review and meta-regression analysis of social functioning correlates in autism and typical development. *Autism Research*, *12*(2), 152–175. https://doi.org/10.1002/aur.2055

Bowler, D. M., Briskman, J., Gurvidi, N., & Fornells-Ambrojo, M. (2005). Understanding the mind or predicting signal-dependent action? Performance of children with and without autism on analogues of the false-belief task. *Journal of Cognition and Development*, *6*(2), 259–283. https://doi.org/10.1207/s15327647jcd0602_5

Burnside, K., Wright, K., & Poulin-Dubois, D. (2017). Social motivation and implicit theory of mind in children with autism spectrum disorder. *Autism Research*, *10*(11), 1834–1844. https://doi.org/10.1002/aur.1836

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain: A Journal of Neurology*, *125*(8), 1839–1849. https://doi.org/10.1093/brain/awf189

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*(3), 314–325. https://doi.org/10.1006/nimg.2000.0612

Catmur, C., Santiesteban, I., Conway, J. R., Heyes, C., & Bird, G. (2016). Avatars and arrows in the brain. *NeuroImage*, *132*, 8–10. https://doi.org/10.1016/j.neuroimage.2016.02.021

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development*, *15*(4), 481–498. https://doi.org/10.1016/S0885-2014(01)00037-5

Cole, G. G., & Millett, A. C. (2019). The closing of the theory of mind: A critique of perspective-taking. *Psychonomic Bulletin & Review*, *26*(6), 1787–1802. https://doi.org/10.3758/s13423-019-01657-y

Conway, J. R., & Bird, G. (2018). Conceptualizing degrees of theory of mind. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), 1408–1410. https://doi.org/10.1073/pnas.1722396115

Conway, J. R., Catmur, C., & Bird, G. (2019). Understanding individual differences in theory of mind via representation of minds, not mental

states. *Psychonomic Bulletin & Review*, *26*(3), 798–812. https://doi.org/10.3758/s13423-018-1559-x

Conway, J. R., Coll, M. P., Cuve, H. C., Koletsi, S., Bronitt, N., Catmur, C., & Bird, G. (2020). Understanding how minds vary relates to skill in inferring mental states, personality, and intelligence. *Journal of Experimental Psychology: General*, *149*(6), 1032–1047. https://doi.org/10.1037/xge0000704

Conway, J. R., Lee, D., Ojaghi, M., Catmur, C., & Bird, G. (2017). Submentalizing or mentalizing in a Level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 454–465. https://doi.org/10.1037/xhp0000319

Crespi, B., Leach, E., Dinsdale, N., Mokkonen, M., & Hurd, P. (2016). Imagination in human social cognition, autism, and psychotic-affective conditions. *Cognition*, *150*, 181–199. https://doi.org/10.1016/j.cognition.2016.02.001

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J. (2004). Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology*, *40*(2), 271–283. https://doi.org/10.1037/0012-1649.40.2.271

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, *85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D. E., & Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLOS ONE*, *8*(9), Article e75468. https://doi.org/10.1371/journal.pone.0075468

Dumontheil, I., Apperly, I. A., & Blakemore, S.-J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331–338. https://doi.org/10.1111/j.1467-7687.2009.00888.x

Durrleman, S., & Franck, J. (2015). Exploring links between language and cognition in autism spectrum disorders: Complement sentences, false belief, and executive functioning. *Journal of Communication Disorders*, *54*, 15–31. https://doi.org/10.1016/j.jcomdis.2014.12.001

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

Edey, R., Cook, J., Brewer, R., Johnson, M. H., Bird, G., & Press, C. (2016). Interaction takes two: Typical adults exhibit mind-blindness towards those with autism spectrum disorder. *Journal of Abnormal Psychology*, *125*(7), 879–885. https://doi.org/10.1037/abn0000199

Filip, A., Białek, A., & Białecka-Pikul, M. (2023). Both syntactic and pragmatic sentence adequacy matters for recursive theory of mind in 5-year-olds. *Cognitive Development*, *66*, Article 101297. https://doi.org/10.1016/j.cogdev.2023.101297

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(34), E5072–E5081. https://doi.org/10.1073/pnas.1610344113

Fletcher-Watson, S., McConnell, F., Manola, E., & McConachie, H. (2014). Interventions based on the theory of mind cognitive model for autism spectrum disorder (ASD). *Cochrane Database of Systematic Reviews*, *2014*(3), Article CD008785. https://doi.org/10.1002/14651858.CD008785.pub2

Frith, U., Happé, F., & Siddons, F. (1994). Autism and theory of mind in everyday life. *Social Development*, *3*(2), 108–124. https://doi.org/10.1111/j.1467-9507.1994.tb00031.x

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder: Autism. *Trends in Neurosciences*, *14*(10), 433–438. https://doi.org/10.1016/0166-2236(91)90041-R

Gao, S., Wang, X., & Su, Y. (2023). Examining whether adults with autism spectrum disorder encounter multiple problems in theory of mind: A study based on meta-analysis. *Psychonomic Bulletin & Review*, *30*(5), 1740–1758. https://doi.org/10.3758/s13423-023-02280-8

Garfield, J. L., Peterson, C. C., & Perry, T. (2001). Social cognition, language acquisition and the development of the theory of mind. *Mind & Language*, *16*(5), 494–541. https://doi.org/10.1111/1468-0017.00180

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, *7*(1), 102–118. https://doi.org/10.1037/arc0000067

Goodman, R. (1989). Infantile autism: A syndrome of multiple primary deficits? *Journal of Autism and Developmental Disorders*, *19*(3), 409–424. https://doi.org/10.1007/BF02212939

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*(1), 26–37. https://doi.org/10.2307/1130386

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, *6*(3), 346–359. https://doi.org/10.1111/1467-7687.00289

Happé, F. G. (1994). 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*(2), 129–154. https://doi.org/10.1007/BF02172093

Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, *66*(3), 843–855. https://doi.org/10.2307/1131954

Happé, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36*(1), 5–25. https://doi.org/10.1007/s10803-005-0039-0

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience*, *9*(10), 1218–1220. https://doi.org/10.1038/nn1770

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659. https://doi.org/10.1111/desc.12148

Hill, E., Berthoz, S., & Frith, U. (2004). Brief report: Cognitive processing of own emotions in individuals with autistic spectrum disorder and in their relatives. *Journal of Autism and Developmental Disorders*, *34*(2), 229–235. https://doi.org/10.1023/B:JADD.0000022613.41399.14

Iao, L. S., & Leekam, S. R. (2014). Nonspecificity and theory of mind: New evidence from a nonverbal false-sign task and children with autism spectrum disorders. *Journal of Experimental Child Psychology*, *122*, 1–20. https://doi.org/10.1016/j.jecp.2013.11.017

Kana, R. K., Maximo, J. O., Williams, D. L., Keller, T. A., Schipul, S. E., Cherkassky, V. L., Minshew, N. J., & Just, M. A. (2015). Aberrant functioning of the theory-of-mind network in children and adolescents with autism. *Molecular Autism*, *6*(1), Article 59. https://doi.org/10.1186/s13229-015-0052-x

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Kuhn, G., Vacaityte, I., D'Souza, A. D. C., Millett, A. C., & Cole, G. G. (2018). Mental states modulate gaze following, but not automatically. *Cognition*, *180*, 1–9. https://doi.org/10.1016/j.cognition.2018.05.020

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PLOS ONE*, *14*(3), Article e0213772. https://doi.org/10.1371/journal.pone.0213772

Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind—An overview of current replications and non-replications. *Data in Brief*, *16*, 101–104. https://doi.org/10.1016/j.dib.2017.11.016

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic

replication study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Lecheler, M., Lasser, J., Vaughan, P. W., Leal, J., Ordetx, K., & Bischofberger, M. (2021). A matter of perspective: An exploratory study of a theory of mind autism intervention for adolescents. *Psychological Reports*, *124*(1), 39–53. https://doi.org/10.1177/0033294119898120

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind. *Psychological Review*, *94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Leslie, A. M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition*, *50*(1–3), 211–238. https://doi.org/10.1016/0010-0277(94)90029-9

Leslie, A. M., & Frith, U. (1987). Metarepresentation and autism: How not to lose one's marbles. *Cognition*, *27*(3), 291–294. https://doi.org/10.1016/S0010-0277(87)80014-8

Leslie, A. M., & Frith, U. (1988). Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology*, *6*(4), 315–324. https://doi.org/10.1111/j.2044-835X.1988.tb01104.x

Leslie, A. M., & Frith, U. (1990). Prospects for a cognitive neuropsychology of autism: Hobson's choice. *Psychological Review*, *97*(1), 122–131. https://doi.org/10.1037/0033-295X.97.1.122

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, *43*(3), 225–251. https://doi.org/10.1016/0010-0277(92)90013-8

Lind, S. E., & Bowler, D. M. (2009). Language and theory of mind in autism spectrum disorder: The relationship between complement syntax and false belief task performance. *Journal of Autism and Developmental Disorders*, *39*(6), 929–937. https://doi.org/10.1007/s10803-009-0702-y

Livingston, L. A., Shah, P., White, S. J., & Happé, F. (2021). Further developing the Frith-Happé animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. *Autism Research*, *14*(9), 1905–1912. https://doi.org/10.1002/aur.2575

Long, E. L., Cuve, H. C., Conway, J. R., Catmur, C., & Bird, G. (2022). Novel theory of mind task demonstrates representation of minds in mental state inference. *Scientific Reports*, *12*(1), Article 21133. https://doi.org/10.1038/s41598-022-25490-x

Lopez, B. R., Lincoln, A. J., Ozonoff, S., & Lai, Z. (2005). Examining the relationship between executive functions and restricted, repetitive symptoms of autistic disorder. *Journal of Autism and Developmental Disorders*, *35*(4), 445–460. https://doi.org/10.1007/s10803-005-5035-x

Low, J., Goddard, E., & Melser, J. (2009). Generativity and imagination in autism spectrum disorder: Evidence from individual differences in children's impossible entity drawings. *British Journal of Developmental Psychology*, *27*(2), 425–444. https://doi.org/10.1348/026151008X334728

Marsh, L. E., Pearson, A., Ropar, D., & Hamilton, A. F. C. (2015). Predictive gaze during observation of irrational actions in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders*, *45*(1), 245–261. https://doi.org/10.1007/s10803-014-2215-6

Millett, A. C., D'Souza, A. D. C., & Cole, G. G. (2020). Attribution of vision and knowledge in 'spontaneous perspective taking'. *Psychological Research*, *84*(6), 1758–1765. https://doi.org/10.1007/s00426-019-01179-1

Milton, D. E. M. (2012). On the ontological status of autism: The 'double empathy problem'. *Disability & Society*, *27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Moessnang, C., Baumeister, S., Tillmann, J., Goyard, D., Charman, T., Ambrosino, S., Baron-Cohen, S., Beckmann, C., Bölte, S., Bours, C., Crawley, D., Dell'Acqua, F., Durston, S., Ecker, C., Frouin, V., Hayward, H., Holt, R., Johnson, M., Jones, E., … the EU-AIMS LEAP Group. (2020). Social brain activation during mentalizing in a large autism cohort: The Longitudinal European Autism Project. *Molecular Autism*, *11*(1), Article 17. https://doi.org/10.1186/s13229-020-0317-x

Mottron, L. (2021). A radical change in our autism research strategy is needed: Back to prototypes. *Autism Research*, *14*(10), 2213–2220. https://doi.org/10.1002/aur.2494

Müller, U., Zelazo, P. D., & Imrisek, S. (2005). Executive function and children's understanding of false belief: How specific is the relation? *Cognitive Development*, *20*(2), 173–189. https://doi.org/10.1016/j.cogdev.2004.12.004

Murray, K., Johnston, K., Cunnane, H., Kerr, C., Spain, D., Gillan, N., Hammond, N., Murphy, D., & Happé, F. (2017). A new test of advanced theory of mind: The "Strange Stories Film Task" captures social processing differences in adults with autism spectrum disorders. *Autism Research*, *10*(6), 1120–1132. https://doi.org/10.1002/aur.1744

Nijhof, A. D., Bardi, L., Brass, M., & Wiersema, J. R. (2018). Brain activity for spontaneous and explicit mentalizing in adults with autism spectrum disorder: An fMRI study. *NeuroImage. Clinical*, *18*, 475–484. https://doi.org/10.1016/j.nicl.2018.02.016

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology*, *125*(6), 818–823. https://doi.org/10.1037/abn0000182

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development*, *60*(3), 689–700. https://doi.org/10.2307/1130734

Perner, J., & Leekam, S. (2008). The curious incident of the photo that was accused of being false: Issues of domain specificity in development, autism, and brain imaging. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *61*(1), 76–89. https://doi.org/10.1080/17470210701508756

Peterson, C., Slaughter, V., Moore, C., & Wellman, H. M. (2016). Peer social skills and theory of mind in children with autism, deafness, or typical development. *Developmental Psychology*, *52*(1), 46–57. https://doi.org/10.1037/a0039833

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*(9), 1353–1367. https://doi.org/10.1177/0956797614558717

Pisani, S., Murphy, J., Conway, J., Millgate, E., Catmur, C., & Bird, G. (2021). The relationship between alexithymia and theory of mind: A systematic review. *Neuroscience and Biobehavioral Reviews*, *131*, 497–524. https://doi.org/10.1016/j.neubiorev.2021.09.036

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Regier, D. A., Kuhl, E. A., & Kupfer, D. J. (2013). The DSM-5: Classification and criteria changes. *World Psychiatry*, *12*(2), 92–98. https://doi.org/10.1002/wps.20050

Reindal, L., Nærland, T., Weidle, B., Lydersen, S., Andreassen, O. A., & Sund, A. M. (2023). Structural and pragmatic language impairments in children evaluated for autism spectrum disorder (ASD). *Journal of Autism and Developmental Disorders*, *53*(2), 701–719. https://doi.org/10.1007/s10803-020-04853-1

Reynell, C., & Harris, J. J. (2013). The BOLD signal and neurovascular coupling in autism. *Developmental Cognitive Neuroscience*, *6*, 72–79. https://doi.org/10.1016/j.dcn.2013.07.003

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266. https://doi.org/10.1037/a0018729

Santiesteban, I., Shah, P., White, S., Bird, G., & Heyes, C. (2015). Mentalizing or submentalizing in a communication task? Evidence from autism and a camera control. *Psychonomic Bulletin & Review*, *22*(3), 844–849. https://doi.org/10.3758/s13423-014-0716-0

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15–21. https://doi.org/10.1016/j.copsyc.2017.04.019

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders. *The Neuroscientist*, *18*(2), 108–113. https://doi.org/10.1177/1073858410397208

Senju, A., & Johnson, M. H. (2009a). Atypical eye contact in autism: Models, mechanisms and development. *Neuroscience and Biobehavioral Reviews*, *33*(8), 1204–1214. https://doi.org/10.1016/j.neubiorev.2009.06.001

Senju, A., & Johnson, M. H. (2009b). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127–134. https://doi.org/10.1016/j.tics.2008.11.009

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Shah, P., Catmur, C., & Bird, G. (2017). From heart to mind: Linking interoception, emotion, and theory of mind. *Cortex*, *93*, 220–223. https://doi.org/10.1016/j.cortex.2017.02.010

Shah, P., Gaule, A., Bird, G., & Cook, R. (2013). Robust orienting to protofacial stimuli in autism. *Current Biology*, *23*(24), R1087–R1088. https://doi.org/10.1016/j.cub.2013.10.034

Sifneos, P. E. (1973). The prevalence of 'alexithymic' characteristics in psychosomatic patients. *Psychotherapy and Psychosomatics*, *22*(2–6), 255–262. https://doi.org/10.1159/000286529

Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: An fMRI study. *Social Neuroscience*, *3*(2), 97–112. https://doi.org/10.1080/17470910701577020

Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental Psychology*, *51*(9), 1190–1200. https://doi.org/10.1037/dev0000039

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Spengler, S., Bird, G., & Brass, M. (2010). Hyperimitation of actions is related to reduced understanding of others' minds in autism spectrum conditions. *Biological Psychiatry*, *68*(12), 1148–1155. https://doi.org/10.1016/j.biopsych.2010.09.017

Stone, V. E., & Gerrans, P. (2006). What's domain-specific about theory of mind? *Social Neuroscience*, *1*(3–4), 309–319. https://doi.org/10.1080/17470910601029221

Tager-Flusberg, H. (2003). Exploring the relationship between theory of mind and social-communicative functioning in children with autism. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 197–212). Psychology Press.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Ten Eycke, K. D., & Müller, U. (2015). Brief report: New evidence for a social-specific imagination deficit in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*(1), 213–220. https://doi.org/10.1007/s10803-014-2206-7

Thiébaut, F. I., White, S. J., Walsh, A., Klargaard, S. K., Wu, H.-C., Rees, G., & Burgess, P. W. (2016). Does faux pas detection in adult autism reflect differences in social cognition or decision-making abilities? *Journal of Autism and Developmental Disorders*, *46*(1), 103–112. https://doi.org/10.1007/s10803-015-2551-1

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *69*(10), 1996–2019. https://doi.org/10.1080/17470218.2014.990392

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: A quick and objective test of theory of mind for adults with autism. *Autism Research*, *4*(2), 149–154. https://doi.org/10.1002/aur.174

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin*, *124*(3), 283–307. https://doi.org/10.1037/0033-2909.124.3.283

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition*, *35*(1), 41–68. https://doi.org/10.1016/0010-0277(90)90036-J

Received May 6, 2024
Revision received September 6, 2024
Accepted November 6, 2024 ■