Mindreading is an Asynchronous Joint Activity: The M-A-J-A Account of Theory of Mind performance, and individual differences.

Ian A. Apperly[1]* Rory T. Devine[1] Stephen A. Butterfill[2]

[1]School of Psychology, University of Birmingham, Edgbaston, Birmingham, United Kingdom.

[2]Department of Philosophy, University of Warwick, Coventry, United Kingdom.

*Corresponding author

i.a.apperly@bham.ac.uk

r.t.devine@bham.ac.uk

s.a.butterfill@warwick.ac.uk

Declaration of interests statement: None

Funding statement: Medical Research Council, Grant/Award Number: R/X002896/1

Mindreading is an Asynchronous Joint Activity: The M-A-J-A Account of Theory of Mind performance, and individual differences.

Abstract

Research on mindreading has been dominated by questions about the presence, absence or nature of mindreading concepts or structures, and by paradigms designed to create the most favourable circumstances for demonstrating such abilities. This focus on competence has led to a neglect of questions about performance. Yet without a theory of performance, mindreading concepts and structures are incapable of explaining how we ascribe particular thoughts and feelings to other people, and it is impossible to explain individual differences in mindreading that are persistent, robust, specific, and consequential for social abilities. We reconsider the theoretical foundations of mindreading to develop an account on which competent mindreading requires generating and selecting mental states that can be recognised as plausible and appropriate by other people, and so is essentially a joint social activity. It is an asynchronous joint activity because, once learned, it can be performed alone. The M-A-J-A (Mindreading as Asynchronous Joint Activity) account explains how mindreading serves as a mediator in human social lives, is shaped by social experience, varies according to that experience, and enables social abilities that would not be the same without its mediating role. The M-A-J-A account can explain a swath of existing findings about individual differences in mindreading that are otherwise puzzling. It provides a framework for understanding how and why mindreading abilities might vary across the lifespan, and for developing interpretable and psychometrically robust measures to study this variation.

Keywords: mentalizing; theory of mind; joint action; individual differences; psychometric

Forty years after the first empirical paper testing children's false belief reasoning, research on mindreading is still viewed predominantly through the lens of early development. This focus on *when* such abilities first occur has consequences for both theoretical and empirical enquiry. Developmental theories tend to emphasise basic structural necessities – such as concepts of belief, desire, intention and the like (e.g., Carpendale & Lewis, 2004 ; Doherty & Perner, 2020; Tomasello, 2018, Wellman, 2014). Empirical studies prioritise paradigms that are sensitive to the detection of such concepts (e.g., Scott & Baillargeon, 2017). Of course, researchers have also examined *how* mindreading develops in ontogeny (e.g. Devine & Lecce, 2021; Gopnik & Meltzoff, 1997; Hughes, 2011; Tomasello, 2010), phylogeny (e.g., Krupenye & Call, 2017; Martin & Santos, 2016), and human history (e.g., Heyes, 2019; Moore, 2021), how it varies between individuals and groups (e.g., Hughes & Devine, 2015; Osterhaus & Bosacki, 2022; Yeung et al., 2024; Lillard, 1998), and its cognitive (e.g., Apperly, 2010; Ferguson & Bradford, 2021) and neural basis (e.g., Gilead & Oschner, 2021). However, the lens of early development means that even this research is dominated by questions about the presence, absence or nature of mindreading concepts or structures, and by paradigms designed to create the most favourable of circumstances for demonstrating such abilities. This is critically limiting for both theoretical and empirical work. Theoretically, concepts may be necessary for mindreading, but they are insufficient because it remains unclear how someone with the necessary concepts is ever in a position to use them effectively. Empirically, as we shall review below, there is increasing evidence of individual differences in mindreading that are persistent, robust, specific, and consequential, yet we lack theories or tasks that cast light on the reasons for this variation. After 40 years we still do not know how people ascribe any specific mental state, nor why some people are better at this than others. Our contention is that these basic questions are related and can only be addressed by rethinking, from the foundations up, ideas about what mindreading is and what makes it possible.

In what follows we first introduce our key ideas in outline form. Next, we develop the theoretical basis for our contention, beginning from foundations, and explore its consequences for how mindreading is conceptualised. In the second half of the paper, we show how our approach can address large gaps in current understanding of longitudinal stability in mindreading and individual differences across the lifespan. We end with a framework for the development of new measures. Such measures would be a major improvement on current approaches because they would make transparent what it means for a person to be more or less good at mindreading.

Outline

What does it mean to "be in a position" to use mindreading concepts effectively? This is a question that we will tackle in stages, but let us start by illustrating the essential problem using by far the most widely-adopted paradigm for mindreading: false belief tasks. In one classic form of false belief task (Wimmer & Perner, 1983) Maxi places his chocolate in the blue cupboard. While he is playing outside his mother moves it to the green cupboard. Maxi returns, wanting his chocolate. The critical question is where will he look for his chocolate? Most people over the age of 4 or 5 years agree that he will look in the blue cupboard, because that is where he falsely believes his chocolate is located. But *why* do most people agree? Strictly speaking, Maxi could reasonably think his chocolate is elsewhere - he might have prior experience of his mother's preference for storing food in the green cupboard, he might have a prior agreement with his mother, or he might have unusual beliefs about the behaviour of physical objects. It is only the carefully-crafted pragmatic constraints of the story (and our sensitivity to those constraints) that mean we are in a position to conceive of only one possible correct answer. Such false belief tasks are an essential workhorse of empirical research on mindreading, but they also obscure the fact that the world is not subject to the kinds of pragmatic constraints that govern storytelling and does not typically come neatly curated into correct and incorrect alternatives for what people think or feel. Outside such tightly constrained tasks mindreaders must take on much more of the work of homing in on plausible mental states. Understanding how we do this is an essential and overlooked puzzle about mindreading.

Why are individual differences in mindreading informative for our purposes? Again, we will tackle this in stages. But let us first establish the intuition that there really is variation in mindreading and explain why this might be puzzling. Imagine, for a moment, your social network of adult friends, colleagues, and acquaintances. If it is anything like ours then you might perceive substantial variation in people's social skills. You may also work in an environment that is highly selective for academic ability. If your work environment is anything like ours, then you might still perceive substantial variation in social skills. That is to say, although general cognitive ability may be relevant for social skills, it is clearly not a sufficient explanation for their variation. As we will later describe, this conclusion is also borne out by empirical investigations. The puzzle is that adults, and indeed children from middle childhood onwards, tend to pass tests for key mindreading concepts (e.g., Wellman, Fang, & Peterson, 2011). There are tests of "advanced" mindreading that continue to pose varying degrees of challenge, and children's earlier performance on tests of mindreading concepts predict later success with "advanced" mindreading (e.g., Devine, White, Ensor & Hughes, 2016). However, there is no theory of "advanced mindreading concepts" to explain what is challenging about these tasks. Thus, the origin of individual differences in "advanced" mindreading is a key point of failure for existing concept-focused theories, and a key challenge for empirical research is to devise mindreading tasks that do justice to the intuition that some people appear better at this kind of thing than others.

Foundations

Giving someone a chef's knife does not make them a chef; it may not even make them good at cutting. Someone looking to employ a chef would be pleased to see that they have a chef's knife but would also want to know that they were able to use it successfully. Research on mindreading has been dominated by the first half of the job description, that is, whether the mindreader has the necessary concepts and rules, while neglecting the second half, that is, whether they can they use them successfully. We believe that key puzzles in research on mindreading can be traced to this neglect.

The issue of successfully using mindreading concepts and rules may seem simple, a minor detail even. Yet comparison with linguistic communication suggests otherwise. While much can be learned from the study of words and linguistic rules, multiple additional fields of study (for example, in phonetics and pragmatics) have proved necessary for addressing how such elements of language are used for successful communication. Likewise, understanding mindreading performance may require fundamentally new ideas. To understand performance, we need an account of successful mindreading.

It is commonly supposed that when we mindread, success consists in identifying facts about mental states as accurately as possible. Framed this way, one starts with an agent whose head is full of beliefs, desires, and intentions that interact to cause the agent’s behaviour. These mental states cannot be observed directly, and so the mindreader’s job is to infer what they might be from the agent’s behaviour. Success is determined by identifying as accurately as possible which mental states the agent in fact has. To many readers this will seem so obviously true that it does not need stating, and although it is seldom stated this supposition is the foundation for most psychological research on mindreading. But one does not need to look far for a contrary view. Dennett’s (1988) “intentional stance” theory is widely cited in empirical research on mindreading. Yet in Dennett’s framing mental states are ascribed as an interpretive gloss over a set of behaviours with no commitment to their neurocognitive basis. If behaviour is predicted successfully then the mental state ascriptions are accurate. The sense in which agents “have” mental states is fundamentally different in this framing because the facts about mental states are determined by whether mindreading succeeds or fails (and not the other way around).

While Dennett focussed on successful prediction of behaviour as the standard of success for mindreading, others have proposed widening the standard to include success in attributing responsibility, making sense of behaviours, and modulating behaviours or ‘mindshaping’ (e.g., Zawidzki, 2013). In what follows we offer a partial defence of these insights and build on them to offer a refinement: successful mindreading requires generating and selecting mental states that can be recognised as plausible and appropriate by other people.

This view of successful mindreading need not be based on taking for granted Dennett’s position. One can also support the view by starting from an opposing theory on which facts about mental states do determine whether mindreading is successful, namely Davidson’s (1973, 1985, 1990) extensively developed theory.[1] Inspired by the success of decision theory, and specifically by the possibility of treating decision theory as an elucidation of what preferences are (Jeffrey, 1983),[2] Davidson pointed to a set of normative requirements which specify a structure to which, he claimed, any mindreading target’s mental states must conform (Davidson, 1980, 7, 12).[3] To illustrate, one requirement is that the target’s beliefs be logically consistent; another is that the target and the mindreader agree about which observations provide evidential support for which conclusion. Davidson demonstrated that this structure makes it possible, in principle at least, to discover a person’s mind through observation and communication.[4] Davidson’s theory also entails a clear view about what makes for better or worse mindreading. In the same sense that there are facts about an object’s weight or temperature, so also there are facts about a person’s mental states; and better mindreading is more accurately identifying those mental states.[5]

[1] The theory’s development is presented in a series of papers starting with Davidson (1973). The most detailed statement is Davidson (1990).

[2] Davidson (1990, 297); Davidson (1985).

[3] ‘if we are to derive meaning and belief from evidence concerning what causes someone to hold sentences true, it can only be … because we stipulate a structure.’ (Davidson, 1980, 7)

[4] ‘What makes the task [of interpretation] practicable at all is the structure that the normative character of thought, desire, speech, and action imposes on correct attributions of attitudes to others, and hence on interpretation of their speech and explanations of their actions.’ (Davidson, 1990, 325; Davidson, 1980, 8)

[5] There is occasionally confusion on this point because Davidson famously allows that there is indeterminacy concerning what someone thinks. Specifically, he rejects the view that ‘each belief has a definite object’ (Davidson, 1974, 154). But indeterminacy is consistent with there being facts about people’s mental states. As Davidson argues, ‘[t]he consequent indeterminacy of interpretation is not […] any more significant or troublesome than the fact that weight may be measured in grams or in ounces’ (Davidson, 1980, 6).

But is this view true? Davidson's theory was never supposed to be an account of how people actually read others' minds (Davidson, 1990),[6] but by articulating the requirements-in-principle for accurate identification of mental states it illuminates some challenges of mindreading in practice. As we will see (in the section on *Formal Models*), these challenges also arise for theories which attempt to describe how people actually read others' minds (e.g., Baker et al, 2017). Attention to the challenges will lead us to a different view about what makes for better mindreading.

The first challenge is the problem of infinite observations. Davidson's theory requires observations of 'a potential infinity' of actions to identify any mental states at all (Davidson 1990, 314). In order to ascribe to Maxi the belief that the chocolate is in the blue cupboard the mindreader would need to observe all of the possible consequences of that belief. Without doing so it is impossible to exclude the potential infinity of variously similar but non-identical beliefs (e.g., that he thinks the chocolate is in the blue cupboard every day except Thursdays). By contrast, practical mindreading involves not merely finite~~,~~ but often very few observations, as illustrated by even the simplified case of Maxi discussed above. How can mindreaders coherently attribute any mental states based on few observations? However they do this, it will amount to selecting one or another set of background assumptions to fill in the missing observations imaginatively. Which background assumptions should be adopted? There are indefinitely many possible background assumptions which could be used to fill in missing observations. We have already illustrated this in the case of Maxi: when asked about his actions or beliefs, many people assume, usually implicitly, that Maxi shares their beliefs about the ways physical objects behave, that

[6] 'All we should require of a theory of truth for a speaker is that it be such that, if an interpreter had explicit propositional knowledge of the theory, he would know the truth conditions of utterances of the speaker.' (Davidson, 1990, 312)

he has no expectations about his chocolate moving while absent, and so on. These implicit assumptions fill in for missing observations. But there are, of course, other assumptions people could make which would lead them to predict different actions and to attribute different mental states to Maxi. And these alternative assumptions would not necessarily be wrong. A group of people might be found to rely on different implicit assumptions, and there would be no logical or rational grounds to criticise them. The most we can say is that there is likely to be an advantage in making roughly the same background assumptions as other people with whom you might seek agreement, including Maxi.

This matters for understanding what skilled mindreading performance is. Having the right concepts and rules is not sufficient. Practical considerations will be needed in selecting background assumptions to fill in missing observations. Whether particular background assumptions are good may depend on a mindreader's aims: what works well for prediction may work less well when the aim is to assign blame or to challenge another person's self-understanding. Given the value of agreement with others on which mental states a person has, whether particular background assumptions are good may also depend on which assumptions other mindreaders will make. One consequence is that better mindreading is not a matter of more accurately identifying mental states, because objective criteria for accuracy are not in play. Another consequence is that better mindreaders may be more flexible in shifting background assumptions between contexts, more willing to consider a wider range of possible mental states when confidence in background assumptions is lower, and better at aligning background assumptions with others. In short, they will be good at generating plausible background assumptions and at selecting the most appropriate from among them.

One might attempt to respond to this challenge by suggesting that mindreading need not involve infinite observations at all because it can be anchored in platitudes (Lewis, 1972). To illustrate, one platitude is that being told something entails knowing it. If Maxi is told that his chocolate is in the green cupboard, then we might infer that Maxi knows his chocolate is in the green cupboard. Such platitudes are surely useful in mindreading (Heider, 1958), but we will overestimate their usefulness if we ignore context. Perhaps the speaker has deceptive motives, or is joking, or simply being sarcastic. Perhaps Maxi does not hear or correctly process the message. An indefinite range of contextual factors can render the platitude incorrect. Appeal to platitudes appears to solve the problem of infinite observations only if we ignore the role of context in mindreading.

The problem of irrationality is the second challenge we face in applying Davidson's theory in practical mindreading. Similarly to decision theory, Davidson's theory applies only on condition that people are ideally rational (and, perhaps even less plausibly, mostly truthful). This condition requires not just flawless inferences: Maxi's every belief, desire and intention must at all times bear in exactly the right way on his actions.[7] Failure of this condition would lead to the conclusion that no humans, nor any other finite animals, have any mental states at all—and our bounded rationality implies, of course, that this condition does fail.[8] A model of how

[7] You cannot say that Maxi's mental states violate the theory's axioms: because the theory aims to be an elucidation of what mental states are, if he fails to conform to the axioms there is—so the theory—simply no way of making sense of the idea that he has mental states at all.

[8] Researchers occasionally respond to a related issue in decision theory by rejecting the idea that decision theory specifies what preferences and subjective expectations are. On such a view, those things exist independently of the theory (e.g. Allais, 1979, 548). This amounts to rejecting one application of decision theory (namely, that of characterising preferences and the rest) while endorsing other applications of it. We agree, of course, that the bounds of rationality are a major obstacle to interpreting decision theory as specifying what preferences and subjective expectations are. The problem is to provide an alternative to decision theory (or to Davidson's theory) which can ensure all researchers have a shared understanding of these states. As far as we are aware, this challenge is yet to be met.

people attribute mental states must allow that people are rational only within limits. Davidson suggested that his theory can be saved only by adopting the formal device that it should apply to parts of people's minds within which the rationality assumption could hold.[9] While other strategies are conceivable (for example, practical mindreaders will simply ignore some of the things people do), any strategy involves deciding where exactly to hold on to a rationality assumption. To illustrate, if Maxi's preferences fail to exhibit transitivity we might select between different possibilities about which are his 'true' preferences and which decision reflects a 'mistake'. This selection cannot, of course, be made on the basis of what is rational. What makes for better or worse selections will depend on practical considerations like the mindreader's culture, as there are likely to be advantages in agreeing with others around us on which mental states someone has. This is a second reason why we cannot think of better mindreading as simply being more accurate in identifying mental states. Because there are many ways you could compensate for irrationality, there are many ascriptions of mental states which will all count as equally accurate. But in practice, people around you are unlikely to recognize you as a competent mindreader unless you agree with them about what Maxi's mental states are, and the ability to agree with them is what underwrites mindreading's utility, even when mindreading alone.

To conclude this section, the problems of infinite observations and irrationality complement each other as one concerns a limit on mindreaders (they are finite) and the other a limit on the mindreader's targets (they are imperfectly rational). Both

[9] Compare Davidson (2004, 181): 'if we are going to explain irrationality at all, it seems we must assume that the mind can be partitioned into quasi-independent structures that interact in ways the Plato Principle [according to which there is no internal irrationality] cannot accept or explain.'

problems motivate rejecting the otherwise attractively simple idea that better mindreading is merely a matter of more accurately identifying mental states.[10] Instead, we must recognize that using mindreading concepts effectively involves identifying assumptions that are plausible and selecting those that are appropriate in a given context. What makes assumptions plausible and appropriate is not something internal to mindreading but is a consequence of what seems plausible and appropriate to people, who of course include both mindreaders and mindreading targets. In this respect, mindreading is not entirely unlike public gift-giving. Success requires other people being disposed to recognise that the gift—or mental state ascription—was appropriate. Ultimately, then, an individual is successful at mindreading if they ascribe states, predict behaviours, assign responsibility, and so on in just the way that would occur if they were involved in a joint mindreading activity with others around them.

Developing the theory

As described in the *Outline* section, concepts and rules are unlikely to be sufficient for mindreading because there may be multiple plausible responses even in simple cases. For example, even carefully crafted false belief tasks appear to be a test of whether someone's intuitions, and their confidence in them, go well beyond anything established by mental state concepts, general principles for their application, and the available evidence in this particular situation.

[10] Of course this brief discussion does not exclude other possible responses to the problems. We also acknowledge that there are entirely different views about the conceptual foundations of mindreading—for instance, some philosophers have held that mental states can be perceived (Smith, 2010; McNeill, 2012), which would motivate a different view. What we have shown here is just that the most influential, best developed theory motivates our conclusions about what it is to be in a position to use mindreading concepts effectively.

To see the extent of this challenge we need a richer example than classic false belief tasks, and so we start by developing the problem of mindreading other people’s thoughts and feelings through the example of gift-giving. Tulip, the head of department, is delighted with the tickets to the opera that her staff bought for her 50th birthday. When Bruno, Tulip’s secretary, was tasked with organising the gift a few ideas had sprung to mind, though on reflection some seemed better than others. Although Tulip isn’t known as an opera buff, it felt like the kind of thing she would like, and her partner confirmed that she didn’t already have tickets. Moreover, it’s the kind of thing she would be happy for everyone to know that she likes. Bruno was confident Tulip would have liked a wine-tasting class even more, but highlighting her liking for alcohol felt potentially awkward, especially in an office where some colleagues do not drink.

This example illustrates that beneath the surface of the everyday activity of gift-giving lies considerable complexity. Success requires abductive “best guesses” about potentially appropriate gifts, which take account of the characteristics of the recipient and their context. Selection among these possibilities that present themselves involves consideration of reasonableness, normativity, and reflexive awareness of others, as well as accommodating particular facts (such as whether the recipient already has the present). In short, selecting a gift draws upon a rich web of information, sources of structure, and constraint, it is essentially relational (involving the giver, the receiver, and audience), and reaches both forwards and backwards in time to take account of relevant history and anticipate future consequences. From everyday experience we might add that it is also something that some people seem distinctly better at than others.

Our contention is that gift-giving illustrates key challenges inherent in many instances of mindreading. In common with longstanding views, mindreading involves the generation and selection of candidate thoughts or feelings for the target. In contrast with most existing views, it suggests that the success of mindreading should be judged by the extent to which it accords with “what people would think” is plausible and appropriate. It even implies that someone mindreading alone is nevertheless attempting to reason from the perspective of a group of people, which is a form of collective reasoning (e.g., Bacharach, 2006; Chater et al., 2022; Schelling, 1960). In what follows we aim to show that these insights motivate a new theory, assimilate a wide range of disparate empirical evidence, and generate productive directions for new research on how social understanding varies between individuals and between groups.

Analogy with modal thinking

The challenge of generating useful answers from a large and ill-defined problem space is widely recognised outside of the literature on mindreading. For example, Phillips et al. (2021) addressed the puzzle of “what comes to mind” during modal thinking about what is possible, but potentially not actual. They note that previous literature consistently “…proposes that we generate the ’alternative possibilities’ fundamental to modal cognition by (i) delimiting a task-relevant partition within the vast space of conceivable possibilities, (ii) considering a smaller subset of particular possibilities within the relevant part of that partition, and then (iii)[11] ordering or evaluating them in task-relevant ways to inform our final modal judgments” (p1027). For example, faced with the task of identifying something for dinner after a disabling

[11] The original text was numbered “ii” but context indicates this must have been a typographical error

trip to the dentist (e.g., Morris et al., 2021), people might (i) form a task-relevant partition of "dishes" (ruling out the much larger set of all inedible things, and things that are edible but not dishes), (ii) generate a smaller "consideration set" of possibilities based upon heuristics including decontextualised value and frequency (iii) select from the consideration set a dish that avoids elements that are hard or crunchy and therefore unsuitable after a disabling trip to the dentist. We believe the challenge of generating and selecting possibilities in modal thinking is a model for the challenge of generating and selecting mental state ascriptions in mindreading , but also that mindreading requires distinctive solutions to these challenges that are enlightening about how mindreading is possible and why it might vary between people.

Mapping the analogy to mindreading.

As already noted, it is often assumed that, for someone with the right concepts and rules, mindreading is a matter of inferring from observed behaviour what people think, feel and intend, much as you might infer from texture and rise what is happening inside a loaf. As we noted in *Foundations,* making such inferences in mindreading would require choosing which of indefinitely many possible background assumptions to make and selecting one among many possible ways of repairing irrationality. The range of possible background assumptions and repair strategies therefore generates a puzzle: How are we so remarkably good at having clear intuitions about what someone else is thinking, feeling, or intending, with high confidence and in agreement with others? Any adequate account of mindreading must explain how these difficulties are overcome.[12] In an analogous way to modal

[12] These concerns are a particular instance of a widely-recognised set of challenges about identifying what is relevant in problem spaces that are very large or only imprecisely specified

thinking we propose that having such intuitions depends on (i) “partitioning” with rule-based constraints, (ii) generating a “consideration set” of plausible possibilities by narrowing down the possible range of thoughts, feelings, or intentions another is having, and (iii) selecting the most appropriate answer from the consideration set.

(i) Partitioning. Mindreading affords many opportunities for using rules to partition the indefinitely large space of possible mental states. For example, since *Anna Karenina* was both written and set in late 19th century Russia, Anna could never have views about such things as global warming or smartphones, or anything else that was not known at the time. Equally, Steve Jobs’ koumpounophobia provides a rule-based limit on the space of possible desires that could be ascribed to him. However, while partitioning in this way clearly reduces the possibilities, as for modal reasoning, the space of possible answers will often remain large, leaving a major challenge for steps (ii) and (iii), which cannot be addressed using rules.

(ii) A consideration set of plausible possibilities. Step two in the model of modal reasoning uses heuristics including decontextualised value and frequency as a basis for generating a consideration set of plausible possibilities. These heuristics are unlikely to be helpful for mindreading because they are insensitive to context and reflect only the person’s individual interests and experiences. What is needed is some heuristic basis on which “what comes to mind” could be conditional on context and the collective interests and experiences of people, not individuals.

(e.g., Fodor 2001; Sperber & Wilson, 1987). We are not proposing a solution to these challenges, which are sometimes thought to be intractable. We are proposing that progress can be made by recognising the existence of these challenges, recognising their particular character in relation to mindreading, and using this analysis to make tractable predictions and interpretations of empirical phenomena.

To address this, we take inspiration from analysis of related challenges that arise when people interact for communicating and acting together. An influential suggestion is that interacting people address these challenges by aligning their actions, and the mental representations and neural processes that govern them. For example, during a discourse, participants make rapid decisions about the linguistic forms they are hearing and producing, based on perceptual input that only partially constrains choices. Many accounts suggest that participants solve this problem through alignment. Over repeated conversational turns, discourse participants increasingly converge in their phonology, word selection, syntax, and meaning representations (e.g., Galotti & Frith, 2013; Pickering & Garrod, 2004). So although on a given turn a speaker might have used many possible linguistic forms to express their meaning, they are more likely to repeat ones that have already featured in the discourse. These alignments simplify the tasks of communicators, by making the same (aligned) representations more available than logically possible alternatives.[13] Importantly, interactive alignment does not entail mindreading, or other inferential processes, but is thought instead to depend upon "low-level" priming mechanisms at multiple levels of processing (e.g., Pickering & Garrod, 2004).

Such alignments have been studied at timescales from milliseconds to minutes, focussing on temporary alignments between participants in a discrete interaction that are dispensed with at the end of the interaction. These are the wrong properties for our purposes. Firstly, it is unlikely that the alignment that is possible within a single interaction would be sufficient to bridge the large gap between generic information from scripts and schemas and the ascription of highly specific thoughts and feelings.

[13] Analogous alignment phenomena occur in non-verbal interaction and coordination (e.g., Brahimi et al., 2010; Sebanz et al., 2006).

Secondly, temporary interactive alignment cannot, by definition, support mindreading outside interactions, yet such mindreading is commonplace. We suggest that these limitations could be addressed with a relatively modest extension of theories of interactive alignment. Research on discourse processing finds that comprehension influences different memory systems over different timescales, with some effects only operating within the limited bounds of short-term memory and others drawing upon and influencing long-term memory (e.g., Rayner, Pollatsek, Ashby & Clifton, 2012). It is only a small step to suppose that alignment, too, develops beyond specific interactions, creating longer-term, asynchronous alignment. Such alignment means that in situations where collective reasoning is relevant, individuals are primed, in a context-sensitive manner, to think similar things in similar ways.

In sum, Morris et al. (2021) suggest that responses “come to mind” for modal reasoning according to their frequency and “cached value” derived from the reasoner’s experience. We suggest that responses “come to mind” for mindreading according to their salience derived from the mindreader’s past interactive alignment. We propose that such asynchronous alignment is critical for explaining how people are in a position to generate plausible candidates for what another person might be thinking or feeling.

(iii) Selecting an appropriate answer from the consideration set of plausible mental states.

Models of modal reasoning commonly propose that plausible candidate responses are evaluated according to task-specific constraints (such as needing to avoid hard or crunchy food) to select a response (Phillips et al., 2021). For mindreading the example of gift-buying illustrates the considerations that guide the selection of an appropriate thought or feeling. These include normative rules about what is morally and socially appropriate, and whether the ascription seems reasonable in light of other things we know about what the person thinks, wants and feels, and other aspects of their situation, time, or circumstances. Considerations are at least sometimes reflexive, such that what I think someone else will want should depend, in part, on their knowing that I or others might find out.

Importantly, such consideration of "appropriateness" is essentially social. What it means to pick the most appropriate possible mental state may just be to pick the one to which a group of people – including the mindreader and the target of the mental state ascription - might also agree, given sufficient time to discuss it. This is why an important determiner of mindreading success will be shared criteria of appropriateness, shared norms and therefore shared background and experience with the targets of mindreading and everyone else involved.

Conclusion: Mindreading is an asynchronous joint activity (M-A-J-A[14])

We propose that mindreading involves partitioning, generation of plausible candidate thoughts and feelings for the mindreading target, and selection of the most appropriate from among this set. Mindreading is "asynchronous" insofar as the criteria of plausibility and appropriateness depend upon a background that exists prior to the current mindreading episode. This is true whether the mindreader is a participant or a spectator in a current social interaction, or if they are alone in the dark trying to imagine why their friend has not called, or whether Tolstoy was consistent in his depiction of Karenin's indifference to his wife's mental life.

[14] Informal feedback indicated that some readers may perceive similarity to a widely-recognised political acronym. For the avoidance of doubt, M-A-J-A is pronounced in the same way as the word "major", following the sounds of the words that it denotes.

Mindreading is nevertheless a “joint activity” because the criteria of plausibility and appropriateness are essentially joint criteria. These criteria are not merely socially learned. Rather, they are constituted by the intuitions and principles with which groups of mindreaders will tend to agree, and to which individuals hold themselves in order to be considered rational and reasonable.

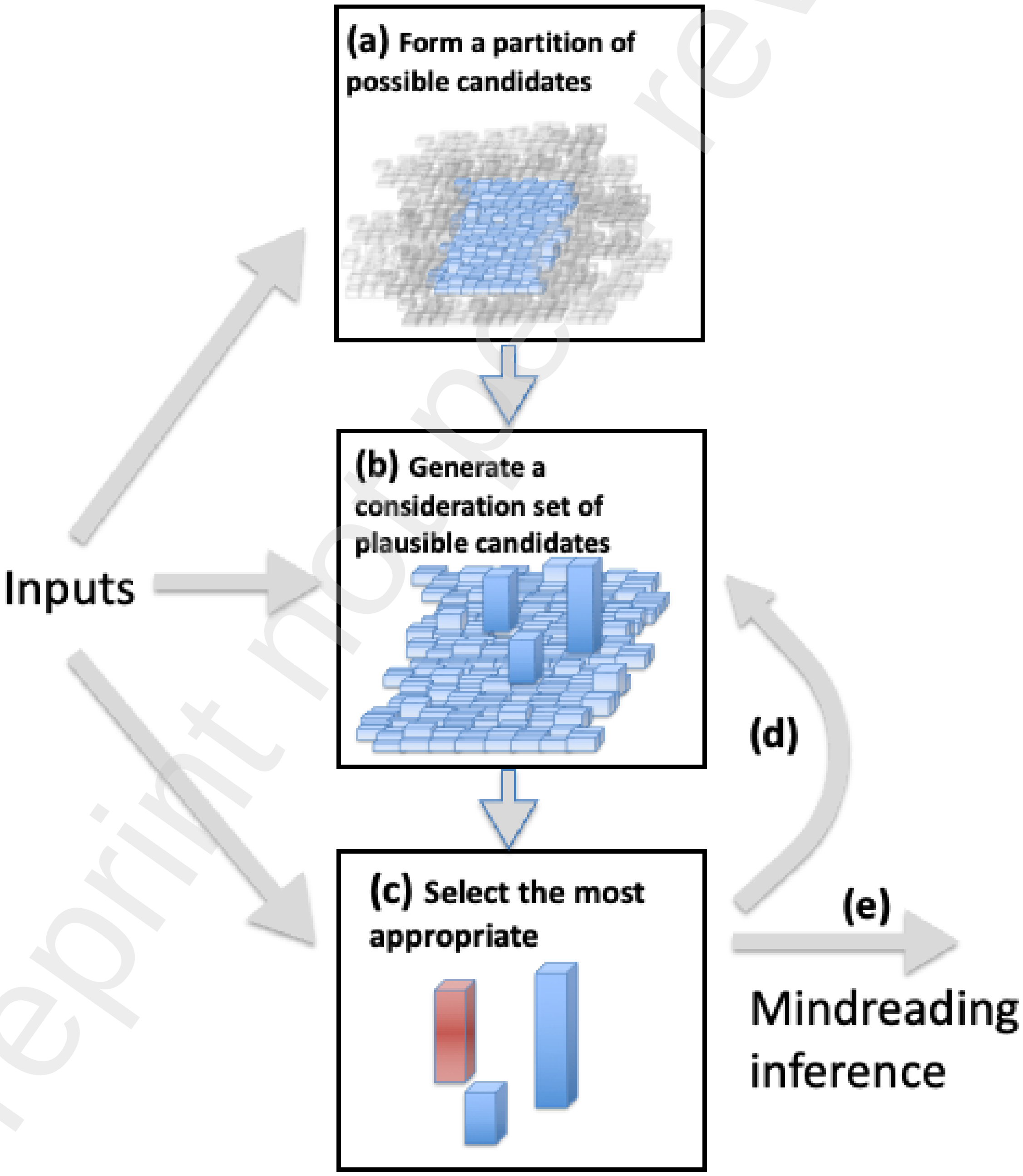
(a) Form a partition of possible candidates
(b) Generate a consideration set of plausible candidates
(c) Select the most appropriate
Inputs
(d)
(e)
Mindreading inference

Figure 1. Schematic model of the process of generation and selection of mindreading inferences. (a) Rule-based constraints form a partition of possible mental states from the field of all available mental states. (b) Plausible candidates are generated by integrating inputs about the target and context with background information that has been structured through interactive social experience. Variation in plausibility is represented by the height of the blue bars. (c) The most appropriate candidate(s) is selected by integrating inputs about the target and context with considerations of normativity, morality, and reasonableness derived from prior social experience. The selected mental state (red bar) may not have been the most plausible at step (b). The selections may become additional inputs to the generation process (d), resulting in further cycles of generation and selection. This ultimately leads to an appropriate candidate being selected as the mindreading inference (e).

Comparison with other accounts

The present account builds upon a rich history of over forty years of theories about the nature of mindreading by philosophers and psychologists. In this section we aim to cast light on the current proposals by identifying crucial points of similarity and contrast with existing accounts. The lens of early development has led many accounts to focus upon the basic structural necessities for mindreading. This leads them to neglect or underestimate the importance of questions about *how* we ascribe mental states that are plausible and appropriate, and severely limits their capacity to explain developmental continuity beyond early childhood or continuing variation in mindreading in adults.

ToMM-SP

A longstanding account proposes that a “Theory of Mind Module” generates possible belief contents while a “Selection Processor” selects between them (e.g., Leslie, Friedman & German, 2004). We do not endorse this modular account, but we gladly acknowledge that our proposal that mindreading involves processes of generation and selection takes inspiration from this and other work that distinguishes between generation and selection during reasoning and decision-making (e.g., Kahneman, 2003; Morris et al., 2021). However, Leslie et al.’s account suffers from two essential problems that relate to the challenges addressed in the present work. First, the ToMM-SP model provides no explanation for how the Theory of Mind Module generates possible belief contents. This should not be surprising since a modular architecture is thought by many to be incompatible with the abductive reasoning required for mental state ascription (Fodor, 2001). Second, Leslie et al.’s account provides no explanation for how the Selection Processor decides which belief to inhibit and which to select (e.g., Apperly, 2010; Doherty, 2008). Leslie, Friedman, and German (2004) hint at the challenge when they suggest that “In reaching its decisions, [the Selection Processor] accesses a learned database of circumstances relevant to selecting between candidate beliefs”. However, because their focus is on the basic structural necessities for mindreading they do not consider the challenges in understanding how such a database develops or how it supports such decisions. From the current perspective the ToMM-SP account addresses much of the challenge of mindreading with a promise that a solution exists somewhere else.

Development of structures and concepts

Whereas the ToMM-SP account assumes that the fundamental structures and concepts for mindreading are innate in a domain-specific module, other influential accounts propose that these are acquired during early development. For example, Perner and colleagues emphasise the acquisition of cognitive structures for metarepresentation (Perner, 1991) or, more recently, for the abstract representation of perspective (Doherty & Perner, 2020). We acknowledge, of course, that a full account of mindreading requires a theory of the cognitive structures involved. However, such structures are empty vehicles for mindreading; necessary for representing mental states but providing no account of how the mindreader comes to ascribe any particular mental state to anyone. In a related programme of work, Wellman and Gopnik (e.g., Gopnik & Wellman, 1992) have led the way in studying children’s acquisition of fundamental mindreading concepts, of perception, knowledge, belief, desire, intention and the like (e.g., Gopnik & Meltzoff, 1997; Wellman, 2014). On one reading these concepts play an analogous role to the structures emphasised by Perner and colleagues, providing the necessary vehicles for mindreading but no account of their use. On another reading, concepts are intended to specify the rules for the use of mindreading – an approach sometimes referred to as “theory-theory” (Gopnik & Wellman, 1992, Gopnik & Meltzoff, 1997). However, for reasons described above they fall far short of achieving that objective and there are reasons to doubt that they could if they tried. Research on the development of structures and concepts can complement but cannot replace the idea that mindreading is an asynchronous joint activity.

Mindreading as simulation

The essential idea behind simulation accounts is that the mindreader need not have an exhaustively specified set of rules for mindreading and can instead use their own mind as a model with which to simulate the functional processes of the target's mind (e.g., Goldman, 2006; Harris, 1992). For this reason, it has been suggested that simulation might avoid the problems of intractable processing described above (Heal, 1996). Accounts typically appeal to "relevant similarity" in biology and experience which ensures that any one brain may serve as the basis for simulating any other, given appropriately similar starting states. However, accounts never explain how biology or experience underwrites the kinds of similarity that are necessary. For this reason, simulation accounts beg the question which should be at the centre of any account of mindreading: how is it that anyone is ever in the position to mindread?[15]

Scripts, Schemas and Intuitive Models

Research in cognitive and social psychology suggests that representations of events, situations, and people are organised into scripts, schemas, and stereotypes (e.g., Cantor et al., 1982; Fiske & Taylor, 1984; Gilbert, 1998; Schank & Abelson, 1977). These ensure that people in a restaurant (to pick a famous example) will share certain expectations about seating, ordering, eating etc., enabling co-ordination between waiters, sommeliers, and diners. Likewise, there is evidence that people have an intuitive model of the structure of personality traits (a "mindspace"),

[15] Simulation is often thought to play a role in interactive alignment. For example, Garrod and Pickering (2004) suggest that "forward models" generated by one's own processes for speech production also serve a predictive role during speech comprehension. This does not mean that simulation provides an account of mindreading.

the accuracy of which is related to their success on advanced mindreading tasks (e.g., Conway et al., 2019; Long et al., 2022). Each of these knowledge structures offers schematic, generalizable information about a situation, event, or person that surely helps with forming a consideration set of plausible possibilities for what someone is thinking. However, this utility is both problematic and limited. It is problematic because over-reliance on generic information is likely to be a source of systematic bias in mindreading (Spaulding, 2018), just as it is in a wide range of other social judgements (e.g., Fiske, 1993). It is limited because scripts, schemas, stereotypes, and other intuitive models are insufficient to explain how we are in a position to go beyond generic information to make flexible, fine-grained, ad-hoc inferences in a particular instance.

Mindreading Accuracy

The challenge of mindreading has often been framed in terms of the "unobservability" of mental states that are presumed to exist in the head of the target of mindreading (e.g., Goldman, 2006; Gopnik & Wellman, 1992; Johnson, 2000; Leslie., 1987; Whiten, 1996). This has motivated attempts to access the "ground truth" of what mindreading targets think and feel by asking them to report on these mental states, and it motivates the idea that the objective of mindreading is accurate identification of those unobservable mental states (e.g., Long et al., 2022; Long, Catmur & Bird, 2024). These views face challenges both in theory and in practice.

The idea that people have privileged and accurate access to their own thoughts has a long and contentious history (Locke, 1689; Russell, 1917; Stich, 1983; Boghossian, 1989; Dretske, 1994; Shoemaker, 1994; McGeer, 1996; Moran, 2001; Bar-On, 2004). It may even be that such self-reports entail the application of interpretive mindreading to oneself (e.g., Carruthers, 2011). Therefore, it is far from obvious that self-reported mental states are a source of ground-truth against which the accuracy of mindreading can be evaluated. Suppose, however, that self-reported mental states were a potential source of ground-truth. Even then, accounts of mental state ascription suggest that accurate identification of these states by mindreaders is not possible in practice, and so cannot be a criterion for mindreading success (see *Foundations*).

In practice, a long tradition of research on "empathic accuracy" has nonetheless assessed mindreaders' accuracy against targets' self-reports. With origins in research on therapeutic interactions Ickes and colleagues (e.g., Ickes et al., 1986; Ickes, 1993; Marangoni et al., 1995) had targets watch a recording of themselves during a clinical interview or an informal interaction and report post hoc on the thoughts and feelings they were experiencing. However, this programme of work finds only limited evidence for stable individual differences in empathic accuracy when participants attempt to infer those thoughts and feelings (Ickes et al., 2000). Accurate judgement of emotional valence depends upon the target themselves demonstrating high emotional expressivity (Zaki et al., 2008), and empathic accuracy is often better-explained via relational factors such as whether (and how well) the target and perceiver are known to each other, rather than the mindreading abilities of participants (e.g., Zaki et al., 2009).

Mindshaping

Zawidzki (2013) claims that the success of human communication and coordination is founded in the interlocking products of social experience. Social experience shapes people's minds in ways that make them mutually interpretable and entitles individuals to reasonable expectations about others while also making them subject to the same reasonable expectations themselves. While Zawidzki's (2013) objectives are not limited to theorising about mindreading[16] there is a clear relationship with the present account. Alignment is a two-way process, and so our proposal that successful mindreading depends on asynchronous alignment implies that the criteria for successful mindreading serve a regulatory as well as a descriptive function, providing criteria to which the target of mindreading should be holding themselves. Consistent with Zawidzki (2013), according to the M-A-J-A account we not only mindread others but also expect everyone (including ourselves) to be held accountable to mindreading interpretations of their own thoughts, feelings, and behaviour. This shared expectation provides further foundation for the effectiveness of socially agreed interpretive principles.

Formal models of mindreading

There have been recent advances in modelling mindreading using formal methods. For example, Bayesian models conceptualise intuitive understanding of behaviour as a "naive utility calculus" according to which people act rationally to maximise their expected rewards (e.g., Baker et al., 2017; Jara-Ettinger et al., 2016). Within a simple simulated world, such models can predict action given an agent's desires and beliefs and infer the most likely beliefs and desires from observation of behaviour. They yield results that accord impressively well with the judgements of children and adults. However, success for "Bayesian ToM" models has only been demonstrated for simple scenarios that maintain tractable reasoning by creating a

[16] Indeed, one of his objectives is to argue that mindreading is of secondary importance to mindshaping in explaining human social abilities.

highly constrained toy world of possible beliefs and desires, and as noted by Stuhlmuller and Goodman (2014) *"At the algorithmic level, it seems likely that exact inference will not scale to models with realistic state spaces."* This illustrates that the problems of infinite observations and rationality raised in *Foundations* also arise for these newer formal models. Clearly these and analogous formal models (e.g., Veissiere et al., 2019) offer exciting prospects for future advances by specifying the general reasoning principles over mindreading concepts. Stuhlmuller and Goodman suggest that there is an outstanding question of understanding how humans cope with situations that cannot be modelled with an exact inference algorithm. We believe this is directly aligned with the M-A-J-A account's objective of understanding how humans make mindreading inferences that are plausible and appropriate without artificial constraints on the state space of possible mental states.

Mindreading is 3rd person and spectatorial

An influential critique proposes that social abilities essentially involve interaction with others (e.g., Redcay & Schilbach, 2019; Schilbach et al, 2013). This view paints mindreading as non-interactive, "third-person", and "spectatorial", and therefore a mis-characterisation of our social abilities. We suggest that a less polarised perspective may be useful. We take it as a given that many social abilities do not require mindreading. Equally, it is clear that mindreading is something that people do, that it serves valuable functions, and stands in need of explanation. Moreover, while mindreading surely occurs during interactions it just as surely occurs outside of them. Therefore, an account of mindreading cannot depend upon the involvement of the mindreader in a current interaction with another person. There is potential for rapprochement, however, because the account developed here *does* suggest that mindreading is essentially a joint activity, albeit one where the "joint" aspects are often asynchronous because they are the products of past interactions and anticipate future social evaluation.

Mindreading is socially constructed

The M-A-J-A account has clear affinities with a variety of claims about the social construction of mindreading. It is also informative to identify points of divergence. A social constructivist perspective is prominent in developmental research (e.g., Carpendale & Lewis, 2004; Fernyhough, 2008; Meins, 2013; Nelson, 1998; Tomasello, 2018), inspired by theoretical ideas from Vygotsky (1931/1997) and Mead (1934), and drawing upon evidence that early mindreading is influenced by social experience with parents and siblings (for meta-analysis see Devine & Hughes, 2018), and linguistic experience, perhaps especially with communicative and pragmatic aspects of language (e.g., Astingon & Baird, 2005). Guided by the lens of early development these approaches have emphasised the social construction of basic mindreading concepts and set themselves in contrast to nativist accounts (e.g., Leslie et al., 2004) or constructivist accounts that emphasise individual discovery and theory-building (e.g., Gopnik & Meltzoff, 1997; Wellman, 2014). However, our account is agnostic on this question about early development: mindreading concepts could be innate, individually or socially constructed, or indeed the emphasis on the acquisition of mindreading concepts could be misplaced. There still remains the question of how people put these abilities to practical use, and this is the focus of our account. This has not typically been a focus of social constructivist accounts of early development, even if they might naturally be extended for this purpose.

Beyond the focus on early development, researchers from a range of disciplines have suggested that mindreading might be influenced or even constituted by "narrative practice" that is essentially socio-cultural in nature (e.g., Bruner, 1990; Hutto 2012; Nelson. 1998) and might be affected by experience with reading character-rich fiction (Zunshine, 2006). Narratives (whether oral or written) are very rich forms, conveying information about people and situations both directly, and more indirectly in ways analogous to scripts and schemas. Like mindreading, storytelling involves turning a potentially infinite space of possibilities into a finite one. For these general reasons narrative experience is a plausible contributor to mindreading, and it is plausible that good storytellers are also good mindreaders. However, a particularly distinctive idea is that folk psychological narratives "...make explicit mention of how mental states ….. figure in the lives, history, and larger projects of their owners" (Hutto, 2009, p11), and that exposure to such narratives, and experience with generating such narratives of one's own, provides the crucial training for understanding how mental states work together in predictions, explanations, and justifications of words and deeds. If this idea is understood in terms of different narrative forms this would suggest that different forms of folk psychological narrative are relevant inputs for mindreading but are limited because they deal only in generalities in a similar way to scripts and schemas. A yet more productive reading of this idea is that narrative communication may be a critical forum for learning the soft constraints on how mindreading is used to make sense of oneself and others, and to test one's own application of those constraints is aligned with that of the critical arbiters: other people.

"Two-systems" accounts

Two of us (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013) have argued that mindreading might be achieved via two types of process that make complementary trade-offs between flexibility and efficiency, in common with longstanding proposals in many other areas of cognition. The account developed here is wholly independent of the success or otherwise of a two-systems approach to mindreading.

The M-A-J-A accont shares some of the same motivating concerns. Part of the motivation for a two-systems account is that processing “belief-like states” that are simpler than beliefs provides a way of avoiding the potentially intractable processing described earlier, if only for the limited range of cases where belief-like states will suffice. However, our two-systems account has not fully addressed the challenge of explaining how people manage to ascribe full-blown beliefs. To that extent the present proposals could be viewed as addressing a significant omission in the two-systems account. Importantly, however, this is also a significant omission from every alternative theory, meaning that the present critique and proposals are relevant irrespective of one’s preference among current theories of mindreading.

Applying the theory to the challenge of understanding individual differences in mindreading

Thus far our account has been motivated by the observation that existing theories of mindreading fail to explain how mindreading concepts can be used in practice. We now turn to variability in mindreading. The motivating intuitions here are that some people are more socially able than others, that their social abilities may vary between contexts (e.g., home versus the workplace), and that variation in mindreading has something to do with this. These intuitions are validated by impressive advances in measurement of individual differences in mindreading. The need for new theory is motivated by the surprising failure of existing theories to explain how such individual differences are possible. We review core phenomena from recent research and highlight two important phenomena: longitudinal stability, whereby individual differences in mindreading in early childhood are stable over time and predict later social abilities; and the existence of individual differences in mindreading into adulthood. We then explain why these phenomena are poorly accounted for by existing theories, before examining how our new account can help.

Core methods and phenomena

Measurement of individual differences requires mindreading tasks on which performance will vary reliably in a given age range. It has been found that tasks originally devised to test for young children's possession of mindreading concepts - such as the false belief task - can be aggregated into batteries that provide reliable measures of individual differences in young children. Individual differences in performance on false-belief task batteries are psychometrically robust: they are explained by a single latent factor (e.g., Hughes, Devine & Wang, 2018), show test-retest reliability (e.g., Hughes et al., 2000), and exhibit rank-order stability over time (e.g., Devine & Hughes, 2019). They are consequential in that they predict later social adjustment (e.g., Lecce & Devine, 2021). Unsurprisingly, since standard false belief tasks were specifically devised to be sensitive tests of young children's concept possession, these tasks also demonstrate ceiling effects beyond age 6. However, the presence of ceiling effects in false belief task performance (as well as measures of desire understanding or second-order false belief understanding) do not necessarily preclude persisting individual differences in mindreading beyond early childhood. To overcome ceiling effects researchers have devised mindreading tasks that involve more subtle or complicated uses of beliefs, desires, and intentions. For example, the Strange Stories task (White et al., 2008; Happé, 1994) presents children with short stories and the Silent Film task (Devine et al., 2013) presents children with short film excerpts, and tests their understanding of the depicted social scenarios that might involve misunderstandings, lies, and double-bluffs. The correct answer is an interpretation determined by the researchers' own judgement, by the majority view of a reference sample of participants, or some combination of these criteria (Yeung et al., 2023). Such tasks successfully raise the ceiling for performance and measure reliable variance into adolescence (Devine, Kovatchev, Grumley Traynor, Smith & Lee, 2023).

These tasks are not only harder. Numerous studies have demonstrated clear gains in mindreading performance across middle childhood and adolescence (e.g., Devine & Hughes, 2013; Dumontheil et al., 2010; Osterhaus et al., 2016; Meinhardt-Injac et al., 2020). These gains are difficult to explain in terms of "new mindreading concepts" but nonetheless appear specifically social because they are not explained by improved performance in language ability or executive function (e.g., Devine & Hughes, 2016; Lecce et al., 2017).

Alongside evidence of continued growth in mindreading performance, there are marked individual differences in performance even within narrow age ranges (e.g., Devine et al., 2023). Rank-order stability in mindreading performance on a given task challenges the idea that individual differences in performance simply capture temporary differences in children's mastery of mental concepts or differences arising from task performance factors (e.g., Hughes & Devine, 2015). Instead, measures like the false belief task appear to provide an early marker of persistent individual differences in mindreading, measured within a narrow window of sensitivity for a given task (Devine, 2021). Moreover, individual differences in performance exhibit longitudinal stability over time (e.g., Banerjee et al., 2011; Lecce et al., 2024). Longitudinal rank-order stability in performance across different age-appropriate measures (i.e. heterotypic stability) suggests that different mindreading tasks might tap into individual differences in the same underlying trait at different points in development (e.g., Devine et al., 2016).

Individual differences in mindreading are consequential because they are linked with meaningful social outcomes. For example, over-and-above language ability, executive function and social motivation, individual differences in mindreading in middle childhood and adolescence are associated with superior social skills (e.g., Ronchi et al., 2020; Devine & Apperly, 2022; Tamnes et al., 2018). Furthermore, in line with evidence from early childhood about the influence of social experience on young children's mindreading (e.g., Devine & Hughes, 2018), children in classrooms where teachers frequently use mental state language perform better on tests of mindreading than other children (Lecce et al., 2022). In summary, individual differences in mindreading from early childhood to adolescence appear to be persistent, robust, specific, and consequential.

There is evidence that individual differences in mindreading persist beyond adolescence into adulthood. In a recent systematic review Yeung et al. (2024) identified 273 studies that used 75 different measures of mindreading with adults. Yeung et al. also argue that current evidence falls short of meeting strong criteria for reliable and valid measurement, in contrast to what has been observed in children. However, current literature provides reason to be optimistic that meaningful individual differences in mindreading persist beyond adolescence, and therefore have the potential to explain persistent individual differences in social ability.

Explaining longitudinal stability

No existing account can adequately explain evidence that individual differences in mindreading are stable over time. Differences in processing capacity or social motivation have the potential – in principle – to explain why some children may be faster than others to pass mindreading tasks, and why differences in mindreading are persistent over time. However, as noted above, processing capacity turns out to be insufficient to explain why earlier mindreading predicts later mindreading and social abilities because variation in mindreading performance predicts social outcomes over and above performance on measures of executive function and language ability (e.g., Devine et al., 2016). Regarding social motivation, recent work suggests this, too, only partially accounts for individual differences in mindreading and social ability (Devine & Apperly, 2021).

Accounts that focus on the structures and concepts necessary for mindreading fail to explain longitudinal stability because mindreading concepts and structures can only ever serve as enablers of social ability, whereas what is needed is for mindreading to serve as a mediator. To illustrate, let us suppose that social experience and social motivation predict mindreading success because they help children to acquire conceptual grasp of the difference between beliefs, desires, and intentions, or between such mental states and other things such as shadows (e.g., Estes, 1988; Wellman & Estes, 1986), or help the development of representational structures necessary for thinking about mental states (e.g., Doherty & Perner, 2020; Perner, 1991). This would *enable* new ways of thinking about mental states. However, concepts and structures are thought to be universally acquired by late childhood. If all children acquire the same set of concepts and structures for mindreading then concepts and structures cannot explain why mindreading ability continues to vary, or why such variance predicts later social ability.

In contrast the Mindreading as Asynchronous Joint Activity (M-A-J-A) account shows how the effective use of mindreading can continue to vary. This means that it can be a *mediator*, that continues to be affected by social experience, social motivation, and other influences, and has unique new influences on social ability. Put simply, the M-A-J-A account has the right form to explain longitudinal stability in mindreading, whereas conceptual accounts do not.

New predictions.

Providing a viable explanation of longitudinal stability yields new and distinctive predictions regarding the measurement of early mindreading abilities. One common practice is to measure early mindreading according to a “scale” ranging from early-acquired to later-acquired concepts. This accords well with conceptual accounts, and as demonstrated by Wellman and colleagues (e.g., Wellman & Liu, 2004; Wellman, Fang & Peterson, 2011) scales of this kind not only show an age-related increase in the number of concepts present, but also a reliable order of emergence, at least within a given culture. A second common practice measures individual differences on batteries of different instantiations of the same task - most often false belief tasks – within a sensitive age range (e.g., Hughes & Devine, 2015). This accords poorly with conceptual accounts, which have no basis for interpreting degrees of performance on tasks designed to operationalise the same concept (Apperly, 2012), yet as reviewed above, performance on such batteries shows stable and meaningful variability. Whereas these two measurement approaches are often used interchangeably, we propose that variation on such measures should instead be conceptualised on two orthogonal dimensions (see Figure 2). The vertical dimension corresponds to the number of mental state concepts that a child appears to understand at a given point in time. The horizontal dimension corresponds to the number of contexts (operationalised as the number of tasks) in which a child can demonstrate appropriate use of a given concept. To the extent that mindreading is a mediator and not merely an enabler of later abilities a child's score on the horizontal dimension should be a better predictor of later mindreading and later social ability, compared with their score on the vertical dimension.

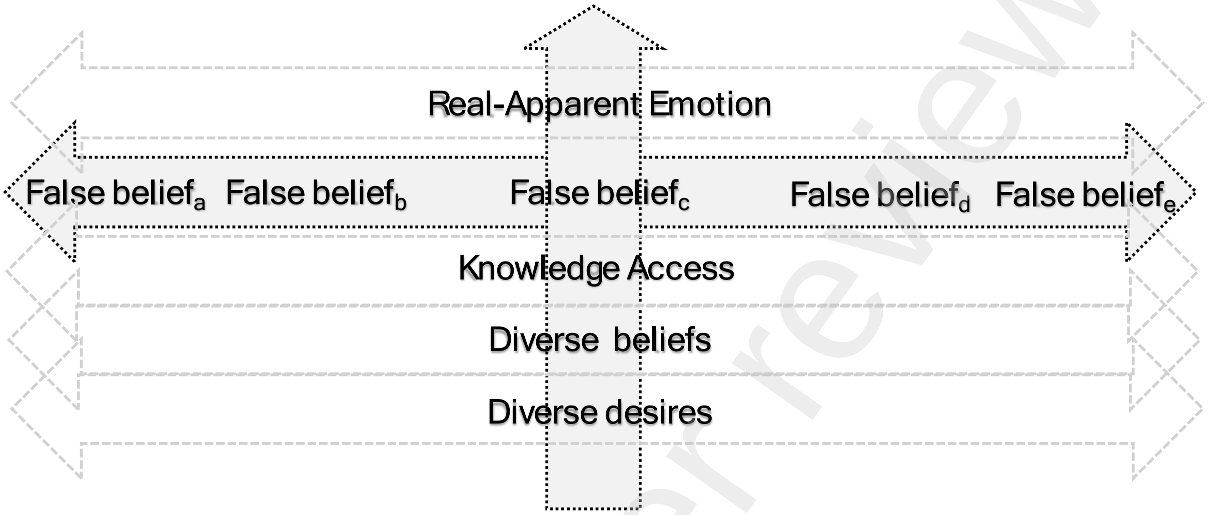

Figure 2. Visualisation of variation in early mindreading on two orthogonal dimensions. The vertical dimension corresponds to the number of concepts currently in a child's repertoire (cf. Wellman & Liu, 2004). The horizontal dimension corresponds to flexible exercise of a given concept across different contexts (False belief a-e). For clarity the latter is illustrated only for the example of false beliefs, but of course any concept can be tested across multiple contexts, as illustrated by the range of tasks that target each concept in the existing literature.

The M-A-J-A account also yields distinctive new predictions about the influence of social experience on mindreading by distinguishing between generation of plausible mental states versus selection of the most appropriate from among them. The account suggests that children's exposure to rich social and communicative interactions in which they become aligned with others will influence their ability to generate plausible possibilities when required to imagine what someone else is thinking or feeling (c.f. Ensor & Hughes, 2008; Meins et al., 2003). It also suggests that children's participation in joint folk psychological activities - of negotiating how mental states feature in reason-giving explanations about other people and oneself - will be a particular influence on their ability to select the most appropriate mental state from among plausible possibilities.

In summary, the M-A-J-A account has the right form to explain longitudinal stability, is readily compatible with evidence of social effects on mindreading over and above cognitive effects and leads to novel predictions and recommendations for improved methods for future work.

Explaining Lifespan individual differences

Since tests of core mindreading concepts are passed during childhood, adolescents and adults cannot be expected to vary in their possession of such core concepts. It has been suggested that adolescents and adults continue to acquire more subtle or sophisticated mindreading concepts, which may not reach a ceiling of sophistication in all adults (e.g., Wellman, 2014). However, no theory describing advanced mindreading concepts has yet been articulated (e.g., Osterhaus & Bosacki, 2022), and it is unclear what new concepts are required for success on existing "advanced" mindreading tasks.[17] At most such tasks typically involve the

[17] For example, Devine and Hughes (2013) describe an item from the "silent films" task as follows: "Harold (the main character) is sitting on the back of a van while he fills in a form. The driver, who is unaware of Harold's presence, is collecting laundry. The driver, who appears to have difficulty hearing, returns to his van, locks the door and drives away. Harold is then trapped in the back of the van." The critical question is: "Why do you think the driver locks Harold in the van?". A correct answer requires participants to acknowledge the critical role of the driver's ignorance: e.g., "Because the driver didn't know Harold was in the van". Many 3-

combination or recursion of core concepts (e.g., an intention thwarted by ignorance; beliefs about beliefs), suggesting that they may require greater processing capacity but not new concepts. New mindreading concepts do not seem necessary to explain lifespan individual differences in mindreading, nor are they likely to be sufficient.

There is evidence from neuropsychology, dual-task interference, and brain stimulation suggesting that adults' successful mindreading depends upon possessing the necessary capacity for working memory and cognitive control (e.g., Apperly, 2010; Frith & Frith, 2012; Gilead & Oschner, 2021; Happe et al., 2017). However, correlations between mindreading performance and tests of working memory and cognitive control are typically modest and are observed inconsistently (e.g., Qureshi et al., 2019; Ryskin et al., 2015), suggesting that these capacities, while necessary, are not critically limiting for many adults on many tasks. A potential exception is highly recursive mindreading, which plausibly makes considerable demands on working memory. However, recent evidence suggests that previous work has significantly overestimated adults' recursive mindreading capacity (Wilson et al., 2023), suggesting that it is unlikely to be a major source of individual differences. Overall, this limited role for processing capacity in explaining individual differences accords with the intuition that general cognitive ability is relevant but insufficient to explain variation in social ability.

year-olds succeed on developmentally sensitive tests of the necessary core concepts of understanding knowledge/ignorance (e.g., Wellman & Liu, 2004), yet there was significant variance in the success of 8- to 13-year-olds in Devine and Hughes' (2013) study. Many participants provided only a partial explanation (e.g., in terms of the driver's desire to continue his rounds), or a completely incorrect explanation (e.g., ascribing a desire to kidnap Harold). We think it unlikely that these participants lacked a concept of knowledge versus ignorance or any other mindreading concept. Instead, we suggest that their difficulty was with identifying the most plausible and appropriate explanation.

Finally, it seems very plausible that variation in social motivation is a source of variation in both social ability in general and mindreading in particular, and there is a small but growing body of evidence consistent with this possibility (e.g., Contreras-Huerta et al., 2020; Pomareda, Devine & Apperly, 2024a). However, without an account of how mindreading can vary in adults, there is no way of explaining how it could be affected by variation in social motivation.

In sum, despite much work, the current literature has made little progress beyond the unsatisfactory conclusion that adults who score higher on tests of mindreading are demonstrating greater mindreading abilities. The M-A-J-A account addresses this challenge by providing criteria for judging the quality of mindreading.

What counts as a “good” mindreading answer?

According to the M-A-J-A account, good mindreading is about agreement with others around us, including - but not only - the target of mindreading. To a first approximation a good interpretation is just what a group of people would jointly conclude given adequate time to discuss it. Discussion establishes consensus on relevant information about the target and their situation, airs potential interpretations, and supports the selection of the best interpretation in light of moral and normative rules, context, and considerations of “reasonableness”. By the end all discussants should acknowledge the quality of the agreed-upon interpretation, even if it was not the one they would have reached independently. It follows that a good mindreader is someone who can simulate the results of such a group process.[18] As described

[18] The idea that mindreaders might simulate the discussion itself fits with the spirit of some neo-Vygotskian accounts that give internalised dialogue a key role in higher mental functions (e.g., Fernyhough, 2007). Our proposal does not entail simulation of the discussion, just the ability to simulate the results of such a discussion.

earlier, the M-A-J-A account suggests that such simulation is dependent upon prior experience of interactive alignment that equips the mindreader with the intuitions to generate a plausible set of candidate mental states, and the shared norms and principles used to select an appropriate mental state from that set.

The first approximation assumes that both the group of mindreaders and the mindreading target are drawn from a similar population, so that it is possible to establish consensus on what information is relevant, what interpretations are plausible, and which is most appropriate. However, it is clearly possible that differences in knowledge, judgements of relevance, and what is moral, reasonable, or normatively correct, could result in groups with different compositions coming to different agreements, or to those agreements being invalid for the target. It follows that a mindreader who is flexible enough to simulate the interpretations of a range of groups and targets will be more successful than one who cannot. With these conclusions it is possible to address the challenge of measuring individual differences in mindreading.

Measurement of individual differences across the lifespan: mindreading interpretations

The M-A-J-A account takes the need for interpretation to be a central feature of mindreading, rather than an unfortunate bug in existing attempts to measure it (c.f. Long, Catmur & Bird, 2024). A first step forward would be for researchers to recognise interpretation as a potential source of variation. Many existing tests of mindreading in adults score responses in ways that confound the quality of participants' responses (i.e. their plausibility and appropriateness) with the mere quantity of mental states mentioned in the response. Using a coding scheme that separated these variables Pomareda et al. (2024b) found independent variation in quality and quantity of mental state descriptions in participants' mindreading explanations, suggesting that these are separate constructs with potentially distinctive roles in predicting social ability.

Second, our analysis highlights the social character of criteria for determining a good interpretation. This should make researchers think carefully about whose interpretations are relevant and valid, who gets to decide which interpretations are better, and whether the answers to these questions are appropriate for a given research question. Suppose, for illustration, that some mindreading stories were created and scored by a straight, white, middle-class, middle-aged, British male academic (the first author acknowledges these characteristics and also owns a chef's knife). We contend that these stories and scoring decisions are unlikely to reflect all the ways in which a broader range of people use mindreading to understand themselves and others, and that this measure may turn out to be easier for people who are more like the researcher than people who are less like him. We suspect that few researchers would disagree. Yet frustratingly, this remains a contention because there is remarkably little evidence on variation in mindreading interpretations, or on the measurement fairness of mindreading tasks across different populations (e.g., Devine & Hughes, 2013; Hughes, Devine & Wang, 2018). This reflects the limits of current theories, which do not explain why these considerations might be important, and it reflects limited use of standard psychometric considerations, (e.g., about the "fairness" of measurement between different groups) when designing and analysing studies of mindreading (Yeung et al., 2023).

What solutions are available? In relation to stimuli, it is feasible to generate mindreading stimuli, questions, and answers systematically to reflect the ways that mental states are used in reason-giving explanations by a broad range of people – not just “experts” or “authorities”. For example, in ongoing work, we have collaborated with young adults with diverse demographic characteristics to create personally meaningful narratives involving social situations that entail mindreading. The creators, not the researchers, were allowed authority over the correct mindreading interpretation of their narrative. Such stimuli have higher face validity than narratives created by an individual or homogenous group of researchers. They also open the way for testing whether variation in the way that different people use mindreading in reason-giving explanations is itself a source of challenge for mindreaders, whether the ease with which mindreaders address that challenge varies, and whether this depends on the diversity of their social experience. This would capture the intuition that a person may be a capable mindreader when “at home” in a familiar environment, but less capable in less familiar environments. Moreover, it suggests that a maximally capable mindreader is one who is able to adapt their mindreading interpretations to the largest range of different mindreading targets and social contexts.

In relation to evaluating performance, the M-A-J-A account suggests that the ultimate criterion should be what the (possibly open-ended) group of people relevant to the mindreader would agree was a good mindreading interpretation. This means that there may sometimes be more than one acceptable mindreading interpretation, that acceptability may vary by context, and that different groups of people may agree on different interpretations. Investigating these possibilities Yeung (2024) asked participants to view scenes involving two protagonists (see Figure 3), and to choose between the interpretations that participants had given spontaneously in a prior study. Without additional context participants showed clear preferences for some interpretations over others, though more than one interpretation was often favoured for a given scene. However, these preferences were influenced by additional contextual information, and by participant characteristics (age of participants in this instance). On the M-A-J-A account the more frequent responses are all contenders for being "good" mindreading answers. However, a good mindreader is one who can accommodate the systematic variation in when any of these potentially good responses is plausible and appropriate.

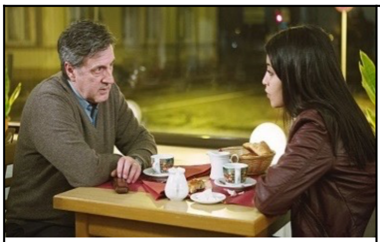 **Response options**	**No context**	**Prime Response (a):** They are meeting because his daughter called him.	**Prime Response (b):** He just invited his daughter out to tell her his decision to divorce her mother.
(a) He is focused on what she is saying.	60%	65%	20%
(b) He is worried about the news he is about to pass on.	20%	20%	80%
(c) He is angry and annoyed about her attitude.	5%	10%	0%
(d) He is feeling disappointed with his daughter.	15%	5%	0%

Figure 3. Example stimuli and illustrative data from Yeung (2024), study 5d. Twenty adult participants per condition judged the response option that best fit the picture. Participants were either given no further context, or prime sentence intended to bias them towards response (a) or (b). Data presented here are from a "young"

adult sample aged 18-25. A further study found that these patterns differed in younger versus older adults (aged 53-60). Whereas younger adults' spontaneous interpretations were most often consistent with option (a), older adults' interpretations were more often consistent with option (c). A person who was sensitive to such variation would be a more capable mindreader.

We want to emphasise that there is no single "correct" approach to these challenges for empirical research. Sometimes it will be appropriate to crowd-source opinion on the correct answers; sometimes authorial intention will be valid; sometimes it will be appropriate to ask a mindreading target what they are thinking or feeling. Design choices should depend both on theory (i.e., the research question) and on empirical considerations (i.e., the scoring criteria to capture individual differences). However, we propose there is potential for researchers to transform the field by using transparent and hypothesis-informed decisions to guide selection and interpretation of existing tasks and creation and scoring of new tasks, and robust psychometric evaluation of how well the tasks serve their research objectives.

Summary and conclusion

To date research on mindreading has failed to address fundamental questions about how we ascribe thoughts and feelings to other people, and why some people seem to be better at this than others. The M-A-J-A account addresses these failures by rethinking what mindreading is and what makes it possible.

We began by challenging the widespread assumption that successful mindreading consists of identifying facts about mental states as accurately as possible. While a target's mental state can be identified accurately in principle (Davidson, 1973; 1990), accurate identification of mental states is a wildly unachievable objective in practice. Instead, part of the art of competent mindreading is generating plausible assumptions and selecting the most appropriate for a given context. What is most appropriate will depend, among other things, on which assumptions others around you would select, and what your aims in mindreading are. A successful mindreader is one who generates and selects those assumptions as if they were making their decisions as part of a group of peers.

How are individuals able to simulate the conclusions of the hypothetical decision-making of a group? In short, because they are socialised. A history of interactive alignment leads members of a population to have similar intuitions about what others might plausibly be thinking or feeling in a particular situation, and to recognise similar principles of normativity, appropriateness, and reasonableness. To the extent that this is true, mindreading is not so much socially constructed as socially constituted as an ongoing but asynchronous joint activity.

The picture that emerges is one that still features mindreading concepts and structures as necessary elements that enable mindreading in-principle. But the practice of mindreading requires the significant additions that we have described here. Besides giving conceptual research on mindreading foundations that can bear its weight, we have shown that our proposals have the capacity to explain extensive empirical phenomena that are currently puzzling and provide a framework for new empirical approaches to studying mindreading beyond the early developmental period that has been the dominant focus for the past 40 years.

References

Abell, F., Happe, 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.

Allais, M. (1979). The So-Called Allais Paradox and Rational Decisions under Uncertainty. In M. Allais & O. Hagen (Eds.), *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of the Decisions under Uncertainty with Allais' Rejoinder* (pp. 437–681). Dordrecht: Springer.

Apperly, I.A. (2010). *Mindreaders: the cognitive basis of "theory of mind".* Hove: Psychology Press / Taylor & Francis Group.

Apperly, I.A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65(5),* 825-839.

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.*

Apperly, I.A. & Wang, J.J. (2021). The Cognitive Basis of Social Interaction in adulthood. In Ferguson, H.J., Brunsdon, V. & Bradford, E. (Eds.) The Cognitive Basis of Social Interaction Across the Lifespan. OUP.

Astington, J. W., & Baird, J. A. (Eds.). (2005). *Why language matters for theory of mind*. Oxford University Press.

Aviezer, H., Trope, Y., & Todorov, A. (2012). Holistic person processing: faces with bodies tell the whole story. *Journal of personality and social psychology*, *103*(1), 20.

Bacharach, M. (2006). *Beyond individual choice*. Princeton: Princeton University Press.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, CD. (2010). Optimally interacting minds. *Science, 329,* 1081–1085.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Banerjee, R., Watling, D., & Caputi, M. (2011). Peer relations and the understanding of faux pas: Longitudinal evidence for bidirectional associations. *Child development*, *82*(6), 1887-1905.

Bar-On, D. (2004). *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Oxford University Press.

Barrett, H. C. (2020). Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends in cognitive sciences*, *24*(8), 620-638.

Boghossian, P. (1989). Content and Self-Knowledge. *Philosophical Topics*, 17(1): 5–26.

Bruner, J. (1990). *Acts of meaning* (Vol. 3). Harvard university press.

Butterfill, S. & Apperly I.A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28(2)* 606-637*.*

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., ... & Singh, S. (2005). Synchrony in the onset of mental-state reasoning: Evidence from five cultures. *Psychological Science*, *16*(5), 378-384.

Cantor, N., Mischel, W., & Schwartz, J. C. (1982). A prototype analysis of psychological situations. Cognitive Psychology, 14, 45-77.

Carpendale, J. I., & Lewis, C. (2004). Constructing an understanding of mind: The development of children's social understanding within social interaction. *Behavioral and brain sciences*, *27*(1), 79-96.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. OUP Oxford.

Chater, N., Zeitoun., H., & Melkonyan, T. (2022). The paradox of social interaction: shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review, 129(3),* 415–437.

Contreras-Huerta, L. S., Pisauro, M. A., & Apps, M. A. (2020). Effort shapes social cognition and behaviour: A neuro-cognitive framework. *Neuroscience & Biobehavioral Reviews*, *118*, 426-439.

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*, 798-812.

Davidson, D. (1973). Radical interpretation. In *Inquiries into truth and interpretation* (pp. 125–139). Oxford: Oxford University Press.

Davidson, D. (1974). Belief and the basis of meaning. In *Inquiries into truth and interpretation* (pp. 141–154). Oxford: Oxford University Press.

Davidson, D. (1980). Toward a unified theory of meaning and action. *Grazer Philosophische Studien*, *11*, 1–12.

Davidson, D. (1985). A new basis for decision theory. *Theory and Decision*, *18*, 87–98.

Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy*, *87*(6), 279–328.

Davidson, D. (1995). Could there be a science of rationality? *International Journal of Philosophical Studies*, *3*(1), 1–16.

Davidson, D. (2004). *Problems of rationality*. Oxford: Clarendon Press.

Devine, R.T. (2021). Individual differences in theory of mind in middle childhood and adolescence. In Devine, R.T. & Lecce, S. (2021) (Eds). *Theory of Mind in Middle Childhood and Adolescence: Integrating Multiple Perspectives* (pp 55-76). Routledge.

Devine, R.T. & Apperly, I.A. (2022) Willing and Able? Theory of Mind, Social Motivation and Social Competence in Middle Childhood and Early Adolescence. *Developmental Science. 25 (1), e13137.*

Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child development*, *84*(3), 989-1003.

Hughes, C., & Devine, R. T. (2015). Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child development perspectives*, *9*(3), 149-153.

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the silent films and strange stories tasks. *Journal of experimental child psychology*, *149*, 23-40.

Devine, R. T., & Hughes, C. (2018). Family correlates of false belief understanding in early childhood: A meta-analysis. *Child development*, *89*(3), 971-987.

Devine, R. T., & Hughes, C. (2019). Let's talk: Parents' mental talk (not mind-mindedness or mindreading capacity) predicts children's false belief understanding. *Child Development*, *90*(4), 1236-1253.

Devine, R. T., Kovatchev, V., Grumley Traynor, I., Smith, P., & Lee, M. (2023). Machine learning and deep learning systems for automated measurement of "advanced" theory of mind: Reliability and validity in children and adolescents. *Psychological Assessment*, *35*(2), 165.

Devine, R. T., & Lecce, S. (Eds.). (2021). *Theory of mind in middle childhood and adolescence: Integrating multiple perspectives*. Routledge.

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental psychology*, *52*(5), 758.

Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and brain sciences*, *11*(3), 495-505.

Doherty, M. J., & Perner, J. (2020). Mental files: Developmental integration of dual naming and theory of mind. *Developmental Review*, *56*, 100909.

Dixson, H. G., Komugabe-Dixson, A. F., Dixson, B. J., & Low, J. (2018). Scaling theory of mind in a small-scale society: A case study from Vanuatu. *Child Development*, *89*(6), 2157-2175.

Doherty, M. (2008). *Theory of mind: How children understand others' thoughts and feelings*. psychology press.

Dretske, F. (1994). “Introspection”, *Proceedings of the Aristotelian Society*. 94(1), 263–278.

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.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., ... & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of autism and developmental disorders*, *36*, 623-636.

Estes, D., Wellman, H. M., & Woolley, J. D. (1989). Children's understanding of mental phenomena. In *Advances in Child Development and Behavior* (Vol. 22, pp. 41-87).

Ferguson, H. J., & Bradford, E. E. (Eds.). (2021). *The cognitive basis of social interaction across the lifespan*. Oxford University Press.

Fernyhough, C. (2008). Getting Vygotskian about theory of mind: Mediation, dialogue, and the development of social understanding. *Developmental Review*, *28*(2), 225-262.

Fiske, S. T., & Taylor, S. E. (1984). Social cognition. New York: Random House.

Fiske, S. T. (1993). Social cognition and social perception. *Annual review of psychology, 44(1),* 155-194.

Fodor, J. A. (2001). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT press.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, *63*, 287-313.

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, *17*(4), 160-165.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T., Fiske, & G. Lindzey (Eds.), The handbook of social psychology (4th ed., pp. 8-150). New York: McGraw Hill.

Gilead, M., & Ochsner, K. N. (Eds.). (2021). *The neural basis of mentalizing*. New York: Springer International Publishing.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Gönültaş, S., Selçuk, B., Slaughter, V., Hunter, J. A., & Ruffman, T. (2020). The capricious nature of theory of mind: Does mental state understanding depend on the characteristics of the target?. *Child Development*, *91*(2), e280-e298.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gopnik, A., & Wellman, A. N. (1992). Why the child’s theory of mind really *is* a theory. *Mind & Language, 7(1&2),* 145-171.

Happé, F., Cook, J. L., & Bird, G. (2017). The structure of social cognition: In (ter) dependence of sociocognitive processes. *Annual review of psychology*, *68*, 243-267.

Harris, P. L. (1992). From simulation to folk psychology: the case for development. *Mind & Language*.

Heal, J. (1996). Simulation, theory, and content. In Carruthers, P., & Smith, P. K. (Eds.). *Theories of theories of mind*. Cambridge university press. 75-89.

Heider, F. (1958). *The psychology of interpersonal relations.* Hillsdale, N.J: Lawrence Erlbaum Associates.

Heyes, C. (2019). Précis of cognitive gadgets: The cultural evolution of thinking. *Behavioral and Brain Sciences*, *42*, e169.

Hughes, C. (2011). *Social understanding and social lives: From toddlerhood through to the transition to school*. Psychology Press.

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry*, *41*(4), 483-490.

Hughes, C., & Devine, R. T. (2015). Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child development perspectives*, *9*(3), 149-153.

Hughes, C., Devine, R. T., & Wang, Z. (2018). Does parental mind-mindedness account for cross-cultural differences in preschoolers' theory of mind?. *Child development*, *89*(4), 1296-1310.

Ensor, R., & Hughes, C. (2008). Content or connectedness? Mother–child talk and early social understanding. *Child development*, *79*(1), 201-216.

Hutto, D. (2009). Folk psychology as narrative practice. *Journal of Consciousness Studies*, *16*(6-7), 9-39.

Hutto, D. D. (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. Cambridge, Mass.: MIT Press.

Jeffrey, R. C. (1983). *The logic of decision, second edition.* Chicago: University of Chicago Press.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. American Psychologist, 58(9), 697–720.

Ickes, W. (1993). Empathic accuracy. *Journal of personality*, *61*(4), 587-610.

Ickes, W., Robertson, E., Tooke, W., & Teng, G. (1986). Naturalistic social cognition: Methodology, assessment, and validation. *Journal of Personality and Social Psychology*, *51*(1), 66.

Ickes, W., Buysse, A. N. N., Pham, H. A. O., Rivers, K., Erickson, J. R., Hancock, M., ... & Gesn, P. R. (2000). On the difficulty of distinguishing “good” and “poor” perceivers: A social relations analysis of empathic accuracy data. *Personal Relationships*, *7*(2), 219-234.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589-604.

Jeffrey, R. C. (1983). *The Logic of Decision, Second Edition*. Chicago: University of Chicago Press.

Johnson, S. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences*, *4*, 22–28.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1503.

Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (2017). Relations between theory of mind and executive function in middle childhood: A short-term longitudinal study. *Journal of experimental child psychology*, *163*, 69-86.

Lecce, S., & Devine, R. T. (2021). Social interaction in early and middle childhood. *The cognitive basis of social interaction across the lifespan*, 47-69.

Lecce, S., Ronchi, L., & Devine, R. T. (2022). Mind what teacher says: Teachers' propensity for mental-state language and children's theory of mind in middle childhood. *Social Development*, *31*(2), 303-318.

Lecce, S., Ronchi, L., & Devine, R. T. (2024). The Effect of Peers' Theory of Mind on Children's Own Theory of Mind development: A Longitudinal Study in Middle Childhood and Early Adolescence. *Developmental Psychology*.

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind.". *Psychological review*, *94*(4), 412.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, *8*(12), 528-533.

Lewis, D. K. (1972). Psychophysical and theoretical identifications. Australasian Journal of Philosophy, 50(3), 249–258.

Lillard, A. (1998). Ethnopsychologies: cultural variations in theories of mind. *Psychological bulletin*, *123*(1), 3.

Locke, J. (1689/1975) *An Essay Concerning Human Understanding*. P.H. Nidditch (ed.), Oxford: Oxford University Press.

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), 21133.

Long, E. L., Catmur, C., & Bird, G. (2024). The Theory of Mind hypothesis of autism: A critical evaluation of the status-quo. *Psychological Review*.

Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of personality and social psychology*, *68*(5), 854.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind?. *Trends in cognitive sciences*, *20*(5), 375-382.

McGeer, V. (1996). “Is ‘Self-Knowledge’ an Empirical Problem? Renegotiating the Space of Philosophical Explanation”. *Journal of Philosophy*, 93(10): 483–515.

McLoughlin, N., & Over, H. (2017). Young children are more likely to spontaneously attribute mental states to members of their own group. *Psychological Science*, *28*(10), 1503-1509.

McNeill, W. E. S. (2012). On Seeing That Someone Is Angry. *European Journal of Philosophy* 20(4), 575–97.

Mead, G. H. (1934)*. Mind, self and society from the standpoint of a social behaviorist. University of Chicago Press.*

Meinhardt-Injac, B., Daum, M. M., & Meinhardt, G. (2020). Theory of mind development from adolescence to adulthood: Testing the two-component model. *British Journal of Developmental Psychology*, *38*(2), 289-303.

Meins, E. (2013). *Security of attachment and the social development of cognition*. Psychology press.

Meins, E., Fernyhough, C., Wainwright, R., Clark-Carter, D., Das Gupta, M., Fradley, E., & Tuckey, M. (2003). Pathways to understanding mind: Construct validity and predictive validity of maternal mind-mindedness. *Child development*, *74*(4), 1194-1211.

Moore, R. (2021). The cultural evolution of mind-modelling. *Synthese*, *199*(1), 1751-1776.

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, NJ: Princeton University Press.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological science*, *32*(11), 1731-1746.

Nelson, K. (1998). *Language in cognitive development: The emergence of the mediated mind*. Cambridge University Press.

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child development*, *87*(6), 1971-1991.

Osterhaus, C., & Bosacki, S. L. (2022). Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, *64*, 101021.

Perez-Zapata, D., Slaughter, V., & Henry, J. D. (2016). Cultural effects on mindreading. *Cognition*, *146*, 410-414.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169-190.Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in cognitive sciences*, *21*(4), 237-249.

Pomareda, C., Devine, R. T., & Apperly, I. A. (2024a). Social Motivation, Over and Above Mindreading, is Associated With Social Support, Autistic Traits, and Depressive Symptoms in Adults. *Manuscript submitted for publication.*

Pomareda, C., Devine, R. T., & Apperly, I. A. (2024b). Mindreading quality versus quantity: A theoretically and empirically motivated two-factor structure for individual differences in adults' mindreading. *Plos one*, *19*(6), e0305270.

Qureshi, A.W., Monk, R.L., Samson, D. & Apperly, I.A. (2020) Does interference between self and other perspectives in Theory of Mind Tasks reflect a common underlying process? Evidence from individual differences in theory of mind and inhibitory control. *Psychonomic Bulletin and Review, 27(1),* 178-190*.*

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495-505.

Ronchi, L., Banerjee, R., & Lecce, S. (2020). Theory of mind and peer relationships: The role of social anxiety. *Social Development*, *29*(2), 478-493.

Russell, B. (1917). Knowledge by Acquaintance and Knowledge by Description. In *Mysticism and Logic*, London: George Allen and Unwin.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*(5), 898.

Schank, R. C., & Abelson, R. P. (1977/2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schelling, T. (1960). *The strategy of conflict.* Harvard University Press.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience 1. *Behavioral and Brain Sciences*, *36*(4), 393-414.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70-76.

Selcuk, B., Gonultas, S., & Ekerim-Akbulut, M. (2023). Development and use of theory of mind in social and cultural context. *Child Development Perspectives*, *17*(1), 39-45.

Shoemaker, S. (1994). Self-Knowledge and ‘Inner Sense’. *Philosophy and Phenomenological Research*, 54(2): 249–314.

Smith, J. (2010). Seeing Other People. *Philosophy and Phenomenological Research* 81(3), 731–48.

Spaulding, S. (2018). *How we understand others: Philosophy and social cognition*. Routledge.

Sperber, D., & Wilson, D. (1987). Precis of relevance: Communication and cognition. *Behavioral and Brain Sciences*, *10*(4), 697-710.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, Mass.: MIT Press.

Stuhlmüller, A., & Goodman, N. D. (2014). Reasoning about reasoning by nested conditioning: Modelling theory of mind with probabilistic programs. *Cognitive Systems Research*, *28*, 80-99.

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, *115*(34), 8491-8498.

Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*, e90.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Wellman, H. M., & Estes, D. (1986). Early understanding of mental entities: A re-examination of childhood realism. *Child development*, 910-923.

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child development*, *82*(3), 780-792.

Whiten, A. (1996). When does smart behaviour-reading become mind-reading? In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 277–292). Cambridge: Cambridge University Press.

Wilson, R., Perez, D. Hruby, A., van der Kleij, S. W. & Apperly I.A. (2023) Is Recursive “Mindreading” Really an Exception to Limitations on Recursive Thinking?. *Journal of Experimental Psychology: General 152*(5), 1454–1468

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.

Vygotsky, L. S. (1997). Genesis of higher mental functions. In R. W. Rieber (Ed.). *The collected works of L. S. Vygotsky (Vol. 4)*. New York: Plenum (Original work published 1931).

Yeung, E. K. L. (2024) Unpublished doctoral thesis. University of Birmingham, UK.

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2023). Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews,* 105481.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological science, 19(4),* 399-404.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9(4),* 478.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Zunshine, L. (2006). *Why we read fiction: Theory of mind and the novel.* Ohio State University Press.