1 Introduction

'Mindreading' refers to the cognitive ability to attribute psychological states to other people. It is distinct from 'social cognition' which has the broader referent base of 'the cognitive structures that facilitate our ability to navigate the social world'. Social cognition is a broader category than mindreading because it is possible to successfully interact with other people without attributing psychological states to them. One might have a perfectly successful social interaction with another by responding to their behaviours, without giving any thought to the psychological states that caused them. Alternatively, one might anticipate the behaviours of another person based on social protocols about how one ought to behave in a particular situation, for example, at a zebra crossing or queuing to get on a bus. Such social protocols extend to how we expect people in different roles to behave, for example, the behaviours we expect of a bus driver and a fellow passenger. These are all instances of interacting with others which do not obviously involve reasoning about their psychological states.

This Element explores the questions of when and why we use mindreading in our social interactions. To do so it must also address the question of what mindreading is. These three questions are naturally interlinked. *What* one takes mindreading to be affects one's views about the situations in which it is necessary for a social interaction. How one construes the purpose of mindreading, that is, our motivations for attributing psychological states to others, similarly affects one's views of which social interactions require it. In the background is the mechanism question, namely *how* do we attribute psychological states to other people? Again, one's position here has knock-on effects for how one addresses the other questions. For example, if one takes mindreading to consist in attributing very basic, non-propositional psychological states to others (an answer to the *what* question), then this could lead to the assumption that the mechanisms that facilitate this ability are also quite basic, requiring relatively few cognitive resources (an answer to the *how* question). This assumption about the mechanisms and constitution of mindreading then shapes answers to the *when* question, leaning towards 'most of the time', because a quick and simple cognitive resource could be in constant use without draining other processing capacities. However, if one scales up the view of what mindreading consists in, for example, maintaining that it is the attribution of more complex psychological states to others such as propositional attitudes,[1] then this in turn shapes both what one thinks about the prevalence of mindreading (is it required for nearly

[1] Of course, one might debate whether propositional attitudes *are* more complex and thus require more cognitive processing than non-propositional states; this possibility is discussed in Section 4.3.

all social interactions, or hardly any?) and one's views on the motivations for mindreading (do we do it in order to have an accurate perception of the others' thoughts or just to gain a rough approximation of what they will likely do next?).

The co-dependence of answers to the *how, what, why* and *when* questions of mindreading means that splitting the questions in this way is, in some sense, artificial. Despite this artifice I think it still provides a useful framework through which to evaluate the large and ever-growing literature on mindreading and social cognition. One of the reasons the literature is so vast is that these topics are of interest to a large array of disciplines: philosophy, neuroscience, social psychology, developmental psychology, anthropology and cognitive ethology – each has its own perspectives and contributions. This Element cannot do justice to even a fraction of these. Instead the aim is to show some of the methods used to tackle these questions, the key points of contention within them and future directions for research in each. In so doing, this Element aims to promote pluralism about mindreading. Until relatively recently philosophers had focused almost exclusively on 'how' mindreading occurs, a debate characterised by the back and forth between 'simulation' and 'theory-theory' approaches which dominated the late 1980s through to the mid-2000s. Yet both simulation and theory-theory approaches agreed upon, and indeed took for granted, answers to the *when, what* and *why* questions. They maintained that mindreading consisted in the ability to attribute propositional attitudes to others (what) in order to explain and predict their behaviours (why) and that it underpinned the vast majority of our social interactions (when). By failing to evaluate and justify their answers to these questions, simulation theorists have assumed a particular characterisation of mindreading which may not necessarily match with reality. If there is no ability that corresponds to that which simulation theorists strive to explain, then their debates about how mindreading proceeds are moot.

The situation for theory-theorists and simulation theorists may not be so bad. We won't know how serious the damage is until we evaluate answers to the *when, what* and *why* questions in more detail. The point is that for too long philosophers (and to some extent psychologists, particularly developmentalists) have assumed that the answers to these questions were agreed upon and that the main puzzle to solve was the 'how' question. This Element aims to show that the story is much more complicated and that previous characterisations of mindreading simplify the phenomenon to the extent that they threaten to warp it out of existence altogether. A comprehensive understanding of mindreading and social cognition needs to confront its complexities to ensure that the phenomenon under study really matches with that which exists in the world.

2 A Brief History of Mindreading

This section offers a short review of how, traditionally, the 'what', 'when', 'why' and 'how' questions of mindreading were addressed in the formative years of the literature, namely the period spanning from the late 1970s through to the mid-2000s. While the theories that emerged and dominated the field during this period are little discussed in contemporary mindreading debates, a retrospective remains useful in order to show (a) the substantial influence of commitments formed during this period in shaping the empirical and theoretical landscape and (b) why the pluralism discussed in Section 3.1 represents such a significant turning point in the progress of the mindreading debates. Most philosophy of mind texts characterise this period as a debate between 'theory-theory' and 'simulation theory', which was a dispute about *how* we attribute psychological states to others. But such a characterisation backgrounds an equally important issue, namely the common ground these competing positions took for granted: why we mindread, when we do so and what it consists in. This review begins with these questions before sketching the simulation/theory debate, thus setting the backdrop for the pluralist revolution discussed in Section 3.1.

2.1 Why: Explanation and Prediction

A key assumption in the early mindreading literature was that the purpose of mindreading was to explain and predict other people's behaviours and that the explanations aimed to fit the deductive-nomological (DN) model associated with Carl Hempel and Paul Oppenheim (1948, Hempel 1965). There were two possible reasons for this focus. First, while opposition to the DN model in philosophy of science was emerging (e.g. van Fraassen, 1980), its popularity as an account of explanation was strong when the early mindreading literature was developing. Second, from an evolutionary perspective it seems that predicting what another will do is more important for generating appropriate responses to them than explaining why they acted as they did. The animal which predicts that another is a threat can engage in avoidance behaviour and thus increase its chances of survival without needing to know anything about *why* the other is acting in that way.[2] Therefore, if one wants an evolutionary explanation for the existence of mindreading, pointing to its function as a predictive device offers a plausible justification for its persistence through the hominid line (and

[2] Of course, knowing why the other acted as they did could enhance survival, e.g. by allowing one to avoid triggering such behaviour in the future. But the point remains that prediction of another's behaviour is both necessary and sufficient for survival in a way that explanation is not.

beyond). These two reasons offer some explanation for the prevalence of the assumptions that (a) we mindread primarily in order to predict other people and (b) that explanations of others' behaviours mimic the form of such predictions, as is captured by the DN model of explanation (Andrews, 2003, 2012). The following passage from an introduction to an edited collection of papers on folk psychology epitomises analytic philosophers' attitudes towards mindreading in the 1980s and 1990s:

> One can explain why I asked my fiancé to marry me by making reference to a host of beliefs and desires I have concerning her and marriage. Moreover one can use these concepts in making predictions about her future behaviour. Knowing that I want to marry my fiancé and that she wants to marry me allows us to predict that, all things being equal, I will say the words 'I do' in a wedding ceremony sometime in the foreseeable future. (Christensen & Turner, 1993, xvi)

The assumption is that the prediction and explanation of the author's behaviour are symmetrical in keeping with the form of a DN explanation:

1. The author wants to be married to his fiancé.
2. The author believes that his fiancé wants to marry him.
3. The author believes that the cultural norms surrounding marriage mean that if she responds 'yes' to his asking her to marry him, she commits to marrying him.
4. *Conclusion:* The author asks his fiancé to marry him.

True to the principles of DN explanation, the form of the explanation is a deductive argument, citing a generalisation about cultural norms (premise 3), and, if one had the explanation prior to the event, one would be able to predict it.

The characterisation of mindreading as a primarily predictive device, with the incidental side effect of generating DN explanations of behaviour, fundamentally shaped the theories that followed, as it entailed that the main puzzles to be resolved were (a) how we attributed psychological states to others and (b) how we acquired knowledge of the rules that governed their interactions. As will be discussed in Section 2.4, this is precisely the task taken on by the theory-theory and simulation accounts. It is also responsible for the huge empirical literature focusing on children's predictions of others' behaviours. The false-belief tasks discussed in Sections 2.2 and 2.4 are premised on the idea that a child's grasp of the concept BELIEF[3] will manifest through their correct

[3] I will use the convention of SMALL CAPS to denote concepts.

predictions of others' behaviours. While there is acknowledgement in some quarters that an ability to predict does not entail an ability to explain (Hood, 2004; Wellman, 2014), the connection between the two is infrequently aired. This silence is perhaps indicative of a tacit assumption that once prediction is in place, explanation naturally follows, as is the case with the DN model of explanation (Andrews, 2012, 37–45).

2.2 What: Propositional Attitudes

Mindreading refers to the process of attributing psychological states to other people. But there is a plethora of psychological phenomena: emotions, thoughts, feelings, knowledge, desires and expectations. Does 'mindreading' cover all of these or just a sub-set?

The focus of the early mindreading literature was almost entirely on how we attribute propositional attitudes to others. One reason for this can be traced to Premack and Woodruff's seminal paper 'Does the chimpanzee have a theory of mind?' in which they argued that their ape, Sarah, demonstrated behaviours which suggested that she attributed propositional attitudes to her human trainers. Sarah was shown short videos, each showing one of her trainers trying to solve a problem (e.g. trying to light the gas heater in her enclosure) and then given three photos, only one of which depicted the solution to the problem (e.g. a lit paper cone, an unlit cone and a burnt-out one). When the video involved her favourite trainer, Sarah succeeded in choosing the correct solution in eleven out of twelve trials; when the video showed her least favourite trainer she chose the correct outcome on only two trials out of eight. Premack and Woodruff maintained that the flexibility of Sarah's behaviour, ranging over diverse scenarios, precluded a behaviourist explanation of her behaviour, arguing instead that Sarah had mindreading abilities. But in a series of responses to the paper, three philosophers observed that Sarah's behaviour was not sufficient to show that she understood that people's behaviours are guided by how they represent the world as being, rather than how the world actually is (Bennet, 1978; Dennett, 1978; Harman, 1978). Each of the tasks required her only to reason about possible states of the world and not how her trainers represented the world. This led the psychologists Hans Wimmer and Josef Perner (1983) to design an experiment to test whether young children understood this aspect of psychological states: the elicited-response false-belief (EFB) task.

There are a number of variations on the EFB task. In 'location-change' tasks, children watch a puppet hide an object, say some chocolate, in one location (a drawer), and leave the scene. In the puppet's absence another character moves

the chocolate from the first location to a new place (e.g. a cupboard). The original puppet returns and children are asked, 'Where will [puppet's name] look for the chocolate?' The result, replicated many hundreds of times, is that two- and three-year-olds routinely answer, 'in location two', with a shift around the children's fourth birthday to the correct response of 'in location one'. 'Contents-switch' tasks (Gopnik & Astington, 1988) are another form of EFB task: a child is shown a tube of Smarties™ and asked what is inside. She responds, 'Smarties!' and the experimenter opens the tube to show that there are pencils inside. The child is then asked what her mummy (waiting outside) will think is in the tube; as before, two-year-olds and early three-year-olds respond, 'Pencils!' with a shift around the fourth birthday to the correct answer ('Smarties'). EFB tasks are so named because they require the child to give a voluntary response to the experimenter's query, either by verbally answering the question or by pointing to where they think the puppet will look. Later sections will examine data documenting variation in the age at which children succeed on EFB tasks, but for now it is sufficient to note that the majority of children tested in Western European and North American settings pass these tasks by the time they are four years old.

It is hard to overstate how much the EFB task has dominated developmental psychology: there are many hundreds of papers on the task, variants and how it relates to other cognitive capacities. The question of what changes in a child's cognition such that she is able to go from failing to passing the task became a central explanandum of mindreading accounts and, in so doing, channelled the focus of those accounts towards explaining how children come to grasp propositional attitudes – specifically the propositional attitude 'belief'. While there are detractors from the standard interpretation that the EFB task tested children's grasp of propositional attitudes (Andrews, 2012; Conway et al., 2020; Gallagher, 2001a), it was, by quite some way, the dominant interpretation of the task.

Some psychologists did manage to widen the focus beyond 'belief' to other propositional attitudes like 'desire' and 'knowledge'. This is evidenced by Henry Wellman's lifetime work on the 'theory of mind scale' (1992, 2014, passim). In the early 1990s, Wellman hypothesised that some psychological concepts were easier to grasp than others and that the development of more complex concepts was dependent on having first acquired the less complex ones (Wellman, 1992).

As can be seen in Figure 1 elicited tasks were used to test children's grasp of these concepts. Toddlers start by passing the 'diverse desires' task at around eighteen months (Repacholi & Gopnik, 1997) and progress through the scale until they pass the 'hidden emotions' task at around five or six years old. While

Task	Brief Description
1. Diverse Desires	Child judges that two persons (the child vs. someone else) have different desires about the same object: Given two possible snacks (carrot, cookie), child states his preference but then must predict snack choice of other person (who has the opposite preference).
2. Diverse Beliefs	Child judges that two persons (the child vs. someone else) have different beliefs about the same object when the child does not know which belief is true or false: Child states her belief that object is in the garage, hears other person's belief that it is in the bushes; child never sees where item is but must predict whether other person will search in the garage or in the bushes.
3. Knowledge-Access	Child judges another person's ignorance about the contents of a container when child knows what is in the container: Child sees toy dog in non-descript drawer, drawer is closed, child judges if other person (who has never seen inside) knows what is in drawer.
4. Contents False Belief	Child judges another person's false belief about what is in a distinctive container when child knows what is in the container: Child sees familiar band-aid box, discovers it has pencils inside, then must judge belief of someone else who has never seen inside.
5. Hidden Emotion	Child judges that a person can feel one thing but display a different emotion: Character is hurtfully teased but doesn't want his friends to know his feelings; child judges how character will feel (sad) and what he will show on his face (happy).

Figure 1 The theory of mind scale (Wellman, 2014, 95)

children differ in the ages at which they pass each task, the order remains almost the same across the world (see Section 5.4 for further discussion). Rhodes and Wellman (2013) showed that preschoolers who pass the knowledge-access tasks can be 'trained' so as to accelerate their performance on false belief tasks (relative to children who did not receive training), which supports Wellman's claim that children must have the earlier concepts in order to grasp the later ones.

The theory of mind scale is more diverse than EFB tasks by studying a wider range of propositional attitude attributions. But the fact remains that its focus is *propositional attitudes*, with emotions making only a token appearance

right at the end.[4] Perhaps, one could speculate, this relates to the evolutionary assumption that prediction is the primary function of mindreading. Correctly attributing propositional attitudes to others can allow one to make specific predictions about what they will do: e.g. If the dominant believes that the fruit is behind the rock, then he will look behind the rock for it. Non-propositional attitudes, perhaps, yield less specific and thus less useful predictions: 'She is angry, so she will engage in generic aggressive behaviours.' This is not to say that recognising emotions is unimportant but rather that such recognition does not facilitate predictions to the degree of specificity exhibited by Sarah. And Sarah, after all, is where this whole debate began. On a related point, emotions and moods may not be characterised by sufficiently robust generalisations so as to be useful within the DN model of explanation, which perhaps indicates another reason for their neglect in this literature.

2.3 When Do We Mindread? Always

> We engage in mindreading for mundane chores, like trying to figure out what the baby wants, what your peers believe about your work, and what your spouse will do if you arrive home late. Mindreading is also implicated in loftier endeavours like trying to glean Descartes's reasons for thinking that many ideas are innate. So pervasive is the role of mindreading in our lives that Jerry Fodor has remarked that if the ordinary person's understanding of the mind should turn out to be seriously mistaken, it would be 'the greatest intellectual catastrophe in the history of our species' (Fodor 1987: xii). (Nichols & Stich, 2003, 2)

The ubiquity of mindreading was often taken for granted by philosophers, as evidenced in the aforementioned passage from Nichols and Stich's book *Mindreading*. In the case of the baby and colleague, the aim is to ascertain the target's psychological state. In the case of the spouse, while the explanatory target is their behaviour, the implication is that ascertaining their psychological state is instrumental to predicting their behaviour (my late arrival will cause *irritation* in my spouse, which in turn causes eye-rolling and sighing behaviours). But do these mindreading-oriented characterisations of mundane interactions match everyday experience? For example, there is a world of difference between having the goal of 'making the baby stop crying' and 'figuring out what the baby wants'. A parent may have learned that a particular toy quiets the baby, handing her the toy without ever contemplating whether it is what the baby *wants* at this moment. A learned behaviour such as this, which can

[4] I am assuming here that emotions need not necessarily be propositional attitudes (Goldie, 2000). Readers who think otherwise can substitute 'emotion' for 'feeling', 'mood' or some other kind of psychological state that is not obviously propositional.

be done non-consciously, does not obviously require mindreading at its execution, although mindreading may have played a part in establishing that learned behaviour (by allowing the parent to establish that the child likes the toy in the first instance). Whether interactions like these necessarily involve mindreading will be questioned further in Section 5.5.

I will refer to the assumption that mindreading is ubiquitous as the 'ubiquity principle'. Stich, Fodor and Nichols are not alone in their acceptance of this principle: many of the most influential voices in the mindreading debates have characterised it in this way (Carruthers, 2020). In so doing, they have driven the explanatory criteria for accounts of mindreading in a particular direction, namely that any good account of mindreading must be compatible with the fact that it drives the vast majority of our social interactions.

2.4 How Do We Mindread? Theory and Simulation

This section reviews the two most well-known accounts of how we attribute psychological states to others: theory-theory and simulation theory. This is likely to be familiar territory to many readers as a comparison of the two is often presented as the gateway into the mindreading debates. Yet, as previously mentioned (Section 2.1), focusing on the differences between the positions backgrounds just how much they share in common. The preceding 'when', 'why' and 'what' sections describe commitments held by both simulation theorists and theory-theorists. Their conflict comes regarding the 'how' question.

Because these views have been well discussed in the literature, they are not described in much detail here.[5] Furthermore, as the positions rest on assumptions which will be challenged through the remainder of this Element, it is not clear how much they can offer to the newly formed social cognition landscape. However, as argued in Section 2.1, it nevertheless remains useful to understand the more established positions in the mindreading debates in order to fully appreciate the motivations of the newer pluralist tradition discussed in Section 3.1.

2.4.1 Theory-Theory

Theory-theory has its heritage in the functionalist approach to mental-state concepts. It refers to that group of views which maintains that psychological state concepts gain their meaning in virtue of their causal-functional roles and that

[5] Readers requiring more information about these accounts can consult Lavelle (2019b) or the main treatises of the positions' advocates, e.g. Carruthers (theory-theory), 2006, 2013; Nichols & Stich (hybrids), 2003; or Goldman (simulation), 2006.

one needs a 'theory of mind' describing these relations in order to grasp them. For example, the concept SEEING is defined by how it relates behaviours, for example, eye direction, to other psychological states (knowing, believing) and new behaviours (e.g. approaching, hiding). What is distinct about the theory-theory view is the claim that at least some of these principles are explicitly represented within the mindreading system; for example, the mindreading system represents the rule SEEING LEADS TO KNOWING (Carruthers, 2011, 2013) and uses it to make inferences about other people's psychological states.

It is important not to conflate the commitment that the principles of the theory of mind are explicitly represented in the mindreading system with the claim that such principles can be introspected. The theory of mind principles that determine our grasp of psychological states are widely considered to be subpersonal; that is, they are principles which a part of our mind uses in order to conceptualise psychological states (Drayson, 2012). As such, they are closed to introspection. Although I can tell immediately that 'I am going to wash my hairs' is grammatically incorrect, I cannot introspect the content of the psychological mechanisms that give rise to that judgement. The only way we can find this out is 'from the outside', with linguists conducting careful experiments and inferring from a collection of my judgements of grammatical correctness which principles may be underlying those judgements. Regardless of whether one takes the theory of mind principles to be explicit or implicitly stored in our cognitive system (Davies & Stone, 2001), the content of those principles cannot be directly introspected.

Within the theory-theory camp is a rift among those who believe that a large part of the theory of mind is innate (Carruthers, 2006, 2011, 2013, passim) and 'constructivists' who claim that there are a few innate rules but the majority of them are learned (Gopnik & Wellman, 2012; Wellman, 2014). As with many philosophical rifts, the disagreement is one of scope rather than content. Henry Wellman maintains that we have innate, domain-specific learning systems which have evolved to be sensitive to specific environmental cues. For example, the system dedicated to learning about mental concepts is sensitive to other people's faces, eyes, intonation, intentional movements and so on. These sensitivities direct the learning system to salient features of the environment from which to acquire information about the domain in question. On Wellman's account, the learning system is Bayesian in structure, using information from the environment to evaluate the probabilities of various possible hypotheses, narrowing down the scope of possible theories by assessing their prior probabilities, likelihood and posterior probability after exposure to a particular learning episode (Wellman, 2014). The Bayesian structure predicts that children learn simpler mental-state concepts (e.g. desire) prior to more complex ones