
# Interpreting signals in other people's behavior to sense things about them and to infer things about their world

| | |
|---|---|
| Journal: | *Social and Personality Psychology Compass* |
| Manuscript ID | SPCO-0976.R1 |
| Wiley - Manuscript type: | Article |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Wu, Wenjie; Lingnan Normal University, Psychology<br>Sheppard, Elizabeth; University of Nottingham School of Psychology, Psychology<br>Mitchell, Peter; University of Nottingham School of Psychology, Psychology |
| Category: | Social Cognition < Compass Sections, Social Psychology and Personality < Subjects, Social Judgment < Social Psychology and personality < Psychology < Subjects, Psychology of Emotion < Social Psychology and personality < Psychology < Subjects, Relationships < Social Psychology and personality < Psychology < Subjects |
| Abstract: | In this article we propose a new framework for investigating how accurately and by what process people read others' minds – a process that requires perceivers to make a retrodictive inference. In this context, we discuss the value of a novel methodological approach that complements the conceptual framework. This framework is formulated on the basis of a series of empirical articles emerging over the past few years in which the ideas appear in nascent form. Retrodiction is the process in which, on observing a person's behavior (often but not exclusively a facial expression), people are equipped not only to sense the underlying inner state but also to infer the event that caused that inner state. Indeed, the goal of mindreading need not always be to identify an inner state explicitly but to infer the event that caused the inner state. Doing so is adaptive in that it permits access to a more expansive view of the world through the lens of another mind. This view of mindreading naturally leads to a reconsideration of methods that are fit for purpose and leads to testable hypotheses. |
| | |

Running Head: *Interpreting signals in behavior*



# Interpreting signals in other people's behavior to sense things about them and to infer things about their world

## Abstract

In this article we propose a new framework for investigating how accurately and by what process people read others' minds – a process that requires perceivers to make a retrodictive inference. In this context, we discuss the value of a novel methodological approach that complements the conceptual framework. This framework is formulated on the basis of a series of empirical articles emerging over the past few years in which the ideas appear in nascent form. Retrodiction is the process in which, on observing a person's behavior (often but not exclusively a facial expression), people are equipped not only to sense the underlying inner state but also to infer the event that caused that inner state. Indeed, the goal of mindreading need not always be to identify an inner state explicitly but to infer the event that caused the inner state. Doing so is adaptive in that it permits access to a more expansive view of the world through the lens of another mind. This view of mindreading naturally leads to a reconsideration of methods that are fit for purpose and leads to testable hypotheses.


*Key words*: mindreading; retrodiction; cognitive development; social cognition; spontaneous behavior; truth condition; response bias; individual differences

# Interpreting signals in other people's behavior to sense things about them and to infer things about their world

## 1. Introduction

Interpreting information signaled in another person's behavior to understand what they think and how they feel, is one of the central aspects of human social cognition. This kind of mindreading is known as retrodiction and refers to the perceiver (a person observing another person's behavior) tracking backwards from the target's (the person being observed by the perceiver) behavior to determine the mental state experienced by the target (Gallese & Goldman, 1998). The goal of this article is to articulate a framework for understanding retrodiction and we shall examine a method that is suited to investigating the underlying process. We shall go on to consider the value of mindreading as an adaptive process that furnishes the mindreader with knowledge of the world that they might not otherwise have known.

## 2. The concept of retrodiction

Although the term was coined by Gallese and Goldman (1998), retrodiction has actually been investigated before that time by Bartsch and Wellman (1989) and by Robinson and Mitchell (1995). In tasks devised by these researchers, young children observed dioramas enacted with dolls in which a protagonist embarked on a futile search owing to his or her false belief. Children aged three and four years were expected to explain the protagonist's behavior by referring to a proximal cause (the

protagonist's false belief, Bartsch & Wellman, 1989) or by referring to a more distal cause (the event that gave rise to false belief, Robinson & Mitchell, 1995). Defined this way, retrodiction refers to the process of identifying the cause of extant behavior, where, according to Teoh et al (2017), the cause can either be understood at a proximal level (an inner state, such as a state of false belief) or at a distal level (a worldly event that gave rise to the inner state, such as the coveted object moving from one place to the other in the protagonist's absence).

Among the various practical benefits of retrodictive inference, one is using another mind as a lens onto aspects of the world that are not perceived directly. Such ability emerges very early in development in a basic form known as 'social referencing' (Sorce, Emde, Campos, & Klinnert, 1985). In real life, people usually have a wide range of information to draw upon in making a retrodictive inference, including the natural and spontaneous facial expression of the target along with other aspects of observable behavior coupled with contextual information. An expressive face can signal an emotional state (Zaki, Bolger, & Ochsner, 2008, 2009), a state that will likely be caused by events in the world (Cassidy, Ropar, Mitchell, & Chapman, 2013, 2015; Kang, Anthoney, & Mitchell, 2017; Pillai, Sheppard, & Mitchell, 2012; Pillai et al., 2014; Sheppard et al., 2016; Teoh et al, 2017; Valanides, Sheppard, & Mitchell, 2017; Wu & Mitchell, 2019). Hence, retrodiction as an account of mindreading assumes that a person's inner states are to some degree signaled in their behavior (e.g., see Gallagher, 2011, for a perspective on how behavior signals inner

states), implying that a person (a perceiver) who is accurate in mindreading is one who is capable of interpreting such signals. Retrodictive inferences not only involve labelling the expression or indeed identifying the target's inner state but also inferring the event (the distal cause) that caused the target's inner state (the proximal cause) that in turn caused the behavior in the target that we currently observe (Teoh et al, 2017). Retrodiction, we argue, is a form of (situational) attribution process, but it is a special kind of attribution that it involves inferring the cause of the event based on perceiving an observable behavioural reaction and via sensing the target's inner state (rather than reasoning from other known facts).

It is noteworthy that being able to retrodict – to make accurate interpretations of a target's behavior -- requires understanding of at least two things: (1) Understanding something about the situation -- certain situations give rise to particular inner states in the target that lead to interpretable signals in the target's behavior; (2) Understanding something about the person – different people emit different signals in relation to a given situation and hence there is no one-to-one mapping between signal and situation. To expand on the latter, the way people (targets) react to events is moderated by who they are, including their emotions, motives, intentions, attitudes, beliefs, knowledge and personality traits (Flavell, Miller & Miller, 1993). Hence, not all targets spontaneously react in the same way to a singular event and target expressions differ from one to another even when they are experiencing the same inner state. That is not to say the relationship between event, inner state and target

expression is completely arbitrary; rather, most spontaneous target expressions will fall within a certain range or category for any given event and inner state. A perceiver who is good at making retrodictive inferences is thus sensitive to the range of reactions, quite often subtle, provoked by a singular event.

## 3. Methodological consideration in research of everyday mindreading--The three principles

A major challenge for research into social cognition is to understand how perceivers fare in real life when events are mercurial and unpredictable, when targets spontaneously react in subtle, unconstrained and occasionally unguarded ways. To understand how perceivers make accurate inferences in real life, we need to investigate how well they interpret subtle and spontaneous target behavior. In view of this, special considerations are required and we outline three principles to guide methodology: (1) the importance of asking perceivers to interpret the *spontaneous* behavior of targets, of the typically subtle kind they might encounter in real life; (2) satisfying the 'truth condition' – the researcher needs to find a way of objectively evaluating a perceiver's interpretation of target behavior; (3) separating accuracy from bias in perceivers' judgments. Each of these principles has been recognized in previous research but perhaps not in combination, although the principles have emerged in combination in recent research on retrodiction. They are abstracted and formulated here to guide future methodological decisions, with the assumption that a

close relationship maintains between theoretical developments in this sphere and relevant methodological considerations.

### 3.1 The importance of asking perceivers to interpret spontaneous target behavior

While some have championed the value of ecologically valid stimuli (Dziobek et al., 2006; Heavey et al., 2000; Zaki & Ochsner, 2011), much research on adults' mindreading has not done so in false belief procedures (Back & Apperly, 2010; Coverse, Lin, Keysar & Epley, 2008; German & Hehman, 2006; Mitchell, Robinson, Isaacs & Nye, 1996), perspective taking tasks (Apperly et al., 2010; Epley, Morewedge & Keysar, 2004; Keysar, Barr, Balin & Brauner, 2000; Keysar, Lin & Barr, 2003), social animation tasks (Abell, Happé & Frith, 2000) and tasks of reading the mind in the eyes (Baron-Cohen, Jolliffe, Mortimore & Robertson, 1997). These tasks usually require the participant to attribute a mental state to a hypothetical target (social agent) or a posed actor based on a fictitious situation, an abstract social animation or based on a static image in the absence of situational context (e.g., photographs of the eyes).

Though these process-based tasks are helpful to probe the cognitive underpinnings of mindreading, they say rather less about how accurately participants routinely make everyday interpretations of other minds when observing subtle and spontaneous behaviour (Zaki & Ochsner, 2011). In the real world we often need to work with fleeting and sometimes unpredictable, rapidly changing stimuli to infer the

cause of the behavior we observe. Making inferences from signals in target behavior is not like reading a novel because these kinds of signals vary from one target to another in reaction to a singular event (within situational limits); unlike natural language, there is no one-to-one correspondence between the signal to be decoded and the event it relates to (Archer & Akert, 1977; Birdwhistell, 1952, 1970).

### 3.2 Satisfying the truth condition

Understanding not just how but *how accurately* perceivers identify targets' inner states, such as what they are thinking or how they feel, is an important question if we assume that the goal of such processing is to understand the social and the physical world in guiding social behavior (Zaki & Ochsner, 2011). According to West and Kenny (2011), the truth condition, thought to be the most essential but elusive component in social perception research, refers to an objective benchmark against which a perceiver's accuracy in interpersonal inferences can be measured. For example, if the perceiver declared that the target is thinking X, how can we verify independently of the perceiver's judgment that the target really is thinking X? Traditional research in a false belief framework achieves this objective by formally defining the mental state of a hypothetical protagonist and then observing whether or not the participant infers that state. But given that we are interested in knowing how accurately perceivers interpret *spontaneous* target behavior of the subtle kind that

prevails in real life, that is amenable to retrodictive inference, satisfying the truth condition presents a challenge.

Would it be a good solution to investigate if a perceiver's inference of the target's inner state matches what the target declares as their inner state? This promising approach seeks to investigate 'empathic accuracy', the ability of one person to read another's mind while the pair are engaged in conversation (e.g., Ickes, Stinson, Bissonnette, & Garcia, 1990; Ickes, 2003, 2009). After the conversation, the two participants watch videos of themselves and write down what they were thinking at any given moment. Subsequently, the participants view videos of their conversational partner and guess what they had been thinking. Judges then compare what targets declared they were thinking with estimations made by perceivers of what the targets were thinking. If there was a fair level of correspondence, the researchers concluded that these conversation partners were to some extent able to read each other's minds. Moreover, one might argue that the truth condition has been satisfied in that we can compare the perceiver's judgment with an 'objective fact' – which is defined as whatever the target declared they were thinking.

On reflection, though, this approach might not offer an ideal solution to the problem. The trouble with this task is that it seems to assume that the target's declaration of their mental state is a reliable and valid source of information. Such an assumption would only be appropriate if it were the case that (1) the target's mind is

transparent to itself – in other words, the target can know what is in their own mind through a process of introspection; (2) the target is sufficiently articulate to be able to translate their inner states into a verbal description and/or adequate vocabulary exists to describe the exact state(s) in question; (3) the target elects to be honest in reporting their inner states. We cannot be sure about any of these three conditions – for example, individuals with alexithymia usually have difficulties verbally describing their own emotions (Grynberg et al., 2012). Hence, there are no grounds for asserting that the truth condition has been satisfied.[1] For this reason, it is useful instead to ask perceivers to retrodict the event that happened in the world that caused the target's inner state that in turn caused their observable behavior. In this case the truth condition is definitely satisfied because we can verify independently whether or not the external event really and truly did happen, though sometimes (but not always, as explained below) we have to sacrifice asking the perceiver explicitly and directly about the target's inner state.



### 3.3 Separating accuracy from bias in the process of mindreading

Fodor (1992) suggests that children's difficulty acknowledging false belief is the product of bias to report their knowledge of the current state of reality. Mitchell et al (1996) also identified such bias in adults' judgments of others' belief states, drawing a parallel with hindsight bias (Fischhoff, 1975) -- and converging evidence was reported by Birch and Bloom (2007). Such bias can occur from underdeveloped or

dysfunctional processes in executive capacities, perhaps preventing the participant inhibiting a prepotent but inappropriate response (Russell, Mauthner, Sharpe, & Tidswell, 1991). Indeed, according to Leslie and Thaiss (1992), bias affecting response selection gives the misleading impression that young children are rather less capable in perceiving others' mental states than is really the case.

When participants are interpreting the behavior of a target, particularly a spontaneous facial expression, they might be biased to interpret the expression negatively (e.g. 'sad') especially if the expression is subtle, as many spontaneous ones are (Kang et al, 2017). According to West and Kenny (2011) it is desirable to identify and isolate bias in order to gain a more accurate impression of ability in mindreading. They define bias as 'any systematic factor that judgments are being attracted toward, besides the truth' (p360). Research would benefit if it were possible to devise a procedure that allows us to investigate the part of performance that is not affected by processing bias (or at least response bias). This follows in the sense that models of ability presume to be founded on data that reveal competence; performance errors, including systematic errors arising from response bias, threaten to mask evidence pertaining to underlying competence.

A widely used procedure in experimental psychology, including research into social cognition, and specifically retrodiction, requires participants to make multiway forced choices, and their responses thus qualify as the dependent variable. If

participants are biased to make one kind of response in preference to another, then it is possible that this bias is not a consequence of deficiency in the capacity being investigated but rather is a nuisance variable that masks underlying ability. Useful procedures for isolating variance in the data associated with response selection bias are signal detection (Macmillan & Creelman, 2005) and Wagner's unbiased hit rate (Wagner, 1993), procedures which compare 'hits' (making a given response when it is correct to do so) with 'false alarms' (making a given response when it is incorrect to do so). Measuring and taking account of false alarms effectively makes adjustments of response base rates, giving rise to a measure that reflects performance after response bias has been partialled out. This principle is highly relevant to research on retrodiction (e.g., Pillai et al., 2012; Teoh et al., 2017; Wu & Mitchell, 2019), as perceivers were asked to select an answer from an array of alternatives in many of the published studies.



## 4. The framework of retrodiction

The framework of retrodiction posits, perhaps controversially, that perceivers capably interpret behavior with consideration of the target's inner state (Bartsch & Wellman, 1989; Gallese & Goldman, 1998; Robinson & Mitchell, 1995) as a preliminary step to inferring the worldly event that caused the target's inner state (Teoh et al, 2017). The process is illustrated in Figure 1.

*Figure 1a, 1b, & 1c about here*

The perceiver, appearing on the left in Figure 1a, observes the expression of the target who is reacting to something that happened in the world, an event that the perceiver can only know by successfully interpreting the target's expression. In this case, the event that happened in the world was the experimenter making a personal call to her boyfriend to discuss dinner (Alkhaldi et al, 2019; Pillai et al, 2012, 2014; Sheppard et al, 2016). While the experimenter was apparently enjoying a scintillating conversation, we might assume the target was irritated and waiting frustratedly to commence the task he had signed up to do. Here, the target's frustration and irritation is signaled in his expression and it seems this expression is readable to the perceiver in as far as the perceiver is able to infer the event in the world that caused the target's expression. Hence, the expression has a proximal cause, which is the inner state being experienced by the target, and this in turn is linked with a distal cause – the event in the world. In the figure, we depict the perceiver's apprehension of the proximal cause – the target's mental state -- as an image of the brain and in this particular case the target's mental state is one of irritation which is represented by the emoticon. We hypothesize that if the target has inferred the distal cause, then they must also have some sense of the proximal cause (the target's mental state).

### 4.1 Alternative hypothesis and explanations

An alternative possibility, though, is that perceivers sense the target's inner state and the event that caused their behavior in parallel (as illustrated in Figure 1b), not in

sequence. If so, then inferring the event in the world that caused the target's behavior is not *dependent upon* sensing the target's inner state in a preliminary step. In principle, this model supposes that the perceiver could infer what caused the target's behavior in the absence of sensing the target's inner state, and if so, then the perceiver's success in retrodictively inferring the external event that caused the target's behavior would not be informative about the perceiver's ability to sense the target's inner states. A more extreme possibility (Figure 1c) supposes that perceivers are behaviorists and do not have any sense of the target's inner state; rather, they only make an inference from the signal in the target's behavior directly to the event in the world that caused the target's behavior.

The models schematized in Figure 1b and Figure 1c are united in implying that perceivers effectively read off an event in the world from the target's behavior (without needing to consider the target's mind). We can never rule out the possibility that perceivers make apparent mentalizing judgments by associating behavior with events that occur in the world (Heyes, 2014). Ultimately, one might argue that a purpose of making inferences from signals in others' behavior is to know something about the world that we would not or could not have known if we lived in a social vacuum and had to rely only on direct experience instead. If perceivers can make such inferences with a degree of accuracy in a context that honors the three principles listed above, then perhaps this in itself should be celebrated, even if sensing inner states plays no part in the process.

Notwithstanding, let us suppose that perceivers do indeed sense that the target has an inner state as a processing step towards inferring the distal cause – the worldly event that provoked the target's reaction. As a thought experiment, imagine that two perceivers, who accurately inferred that the target was in the presence of an experimenter who was rudely using her mobile phone, each articulated what they sensed to be the target's inner state: one sensed that he felt irritated and one sensed that he felt bored. Does this imply that one of them must be wrong? If so does it imply that a perceiver need not sense the target's inner state accurately to make an accurate inference of the distal worldly event that provoked the target's reaction? Surely this would imply that being able to sense the target's inner state is not a necessary processing step for the perceiver in accurately inferring the distal cause of the target's behavior, effectively suggesting that what is depicted in Figure 1a is wrong.

However, we propose that what is depicted in Figure 1a remains the most compelling process account, even though data tell us that different perceivers sometimes do actually impute different inner states on viewing the same target behavior. Sheppard et al (2016) asked perceivers to freely describe a particular instance of target behavior (in the absence of any constraint associated with forced-choice responding) and a multitude of labels emerged, indeed including a variety of inner state terms. But this need not imply that perceivers entertain a loose mapping between the proximal (the particular target inner state) and the distal (the worldly event); rather, it could and probably does imply a loose mapping in perceivers

between what they sense the target is experiencing and the label they select to declare what they believe to be the target's inner state. The key point here is that as suggested in Pillai et al (2012), we assume perceivers sense the target's inner state as a kind of bodily resonance that gives rise to an echo of the target's inner state, enabling the perceiver to share that inner state (Gallese, 2005) -- something Goldman and de Vignemont (2009) call 'low-level mentalizing.'

In contrast, 'high-level mentalizing' in Goldman and de Vignemont's (2005) taxonomy involves explicit propositional representations of the kind that can be verbalized. The cognitive elaboration to enable mapping from low-level to high-level mentalizing probably will depend on what aspect of bodily resonance the perceiver attends to (is it the aspect of irritation or the aspect of boredom?) coupled with their ability to attach labels to feelings. Pillai et al (2012) argued that such explicit elaboration does not illuminate the process of retrodiction and we reinforce that argument here. In short, then, when two perceivers explicitly impute different inner states to a target on viewing a singular sample of behavior, this remains perfectly compatible with the processing route sketched in Figure 1a. Moreover, the possibility that perceivers explicitly impute different inner states in relation to the same target behavior could well be a sign that explicit, high-level propositional mentalizing is not particularly relevant to retrodiction or indeed to a broad range of mindreading activities.

**4.2 Is it possible to investigate retrodiction according to the three principles?**

Reading other minds is something that happens in a social context, it is not something that happens in social isolation. If it did not happen in a social context, then there would not be a mind available to be read. Consequently, a priority is to understand how people interpret the kind of clues one might encounter in a social context, namely states of mind that are signaled in the subtle, fleeting and spontaneous behavior of a target (Zaki & Ochsner, 2011). It follows, therefore, that the aim should be to investigate how perceivers fare in processing such stimuli, as was done in Pillai et al (2012) and the articles that built on that research (*Principle 1, use spontaneously generated stimuli*).

When the perceiver senses that the target is experiencing state X, how can we know that the target is really and truly experiencing state X (West & Kenny, 2011)? Specifically, how can we know this independently of what the target says they are experiencing (because they may not know even if they have the illusion that they do know) and independently of any consensus view offered by panels of judges (because even if there is consensus, it does not follow they are right)? The answer we offer to this apparently insuperable problem is to ask perceivers not about the target's inner state (the retrodictive proximal cause of the target's behavior), but to infer the event that caused the target's inner state (the retrodictive distal cause of the target's behavior). The perceiver's judgment can then be measured against an objective and

independently verifiable fact: When the perceiver declares that the target's reaction was caused by Event X, we (the experimenters) can determine whether or not Event X really and truly did occur (*Principle 2, satisfy the truth condition*).[2]

The price of satisfying the truth condition involves a tradeoff such that perceivers are not actually asked about the target's inner state; instead, they are asked what caused the target's behavior – an event that happened in the world. This procedure is not unique in researching understanding of inner states. Indeed, Wimmer and Perner (1983) took pains to devise a task in which children were deliberately not asked explicitly about the thoughts of a protagonist in case their lack of mental state vocabulary (Johnson & Maratsos, 1977; Johnson & Wellman, 1980) caused them to underperform, potentially giving the wrong impression that they lacked knowledge of inner states when in fact they merely lacked practical knowledge of mental state vocabulary. In their classic false belief task, Wimmer and Perner (1983) did not ask children explicitly where Maxi thought the chocolate was located but rather asked them to predict where Maxi would search. In so far as competent participants understand that behavior (including searching) is determined by inner states, they should judge that Maxi will search according to his belief and not according to the child participant's own belief. By the same token, when participants explain the target's behavior, we assume they do so by retrodictively considering the target's mental state en route to inferring the event in the world that caused the target's mental state (Robinson & Mitchell, 1995).

However, the worldly event that caused the target's inner state could actually be a cue-word, such as 'grateful' (i.e., 'Think of a time you felt grateful' – Valanides et al, 2017; Wu & Mitchell, 2019). On observing a video of the target's behavior, perceivers could be asked to infer the cue word that appeared before the target, or they could even be asked to infer what the target was thinking. We predict in this particular case (where a question about the worldly event and a question about the target's inner state can be correctly answered with the same response) that their accuracy would be no different under these two conditions. Hence, this procedure can indeed be adapted to explicitly ask perceivers to make judgements of other's inner states, even though in many cases of retrodiction, the target's inner state is not identified explicitly by the perceiver.

When perceivers make errors in retrodiction or other forms of mindreading, what does this reveal about the status of their (lack of) ability? Does it tell us they lack the ability to read minds or does it tell us that even though they are capable of reading minds, pervasive bias in their response selection leads to errors, perhaps systematic errors that mask underlying competence (Mitchell et al., 1996)? In a task that meets the condition set out in *Principle 2*, above, bias could result in the perceiver defaulting to one response option over alternatives in a forced choice task of the kind used in retrodictive mindreading procedures. For example, imagine the perceiver's task is to infer whether the target is looking at a negative or positive image (North, Todorov & Osherson, 2010, 2012). In many cases, the target reaction will be subtle, meaning that

targets look somewhat neutral. Neutral expressions may be interpreted by perceivers as negative expressions (Kang et al, 2017), which could result in a preponderance of 'negative' responses. This bias in perceiver judgments needs to be separated from aspects of performance which speak to perceivers' competence or lack of competence (West and Kenny, 2011 – *Principle 3, measure and take into account response bias*). In the kind of example offered here, it is possible to measure and take into account response bias in the kind of forced choice task used in retrodiction research by using either signal detection for coding data (Pillai et al, 2012; Wu et al, 2016, 2017; Wu & Mitchell, 2019) or Wagner's unbiased hit rate (Sheppard et al, 2016), as described above.

**4.3 Sensing inner states – what is the evidence?**

Retrodictive inference enables the perceiver to guess what happened to others from their reactions (Cassidy et al., 2013, 2015; Pillai et al., 2012, 2014; Sheppard et al., 2016), to know the emotion being expressed by a third party that the target accompanies (Kang et al, 2017), to infer what the target was thinking as demonstrated by accurate identification of a cue word presented to the target (Wu & Mitchell, 2019). In the study conducted by Kang et al (2017), for example, targets observed a succession of faces posing expressions displayed on a computer screen. Videos of the targets as they viewed the expressions were subsequently shown to perceivers who were tasked with inferring which emotional expression the target was viewing at any

one time. Despite the very weak signal, perceivers were nevertheless surprisingly accurate in inferring which expression targets were viewing, even when those perceivers were as young as 7 years. Presumably, targets spontaneously and subtly imitated the expressions they were viewing, giving rise to a readable signal.

The task described above involves a retrodictive inference concerning spontaneous behavior – the target spontaneously and subtly imitated the expression they were viewing (Principle 1) -- and it is one that satisfies the truth condition, for we know for a fact which expression the target was really and truly viewing at any given moment (Principle 2). Furthermore, the data could be coded using Wagner's unbiased hit rate (or signal detection), which proved to be essential given that perceivers were biased to judge that targets were looking at a sad expression (Principle 3).

Notwithstanding, accurate performance of the perceivers could be explained reductively and contrary to theory. The theory hypothesizes that accurate perceiver performance is a product of retrodiction that involves sensing the target's mental reaction (their inner state) caused by what they can see (a facial expression) but a rival explanation reductively says that perceivers disregarded the instruction to infer which expression targets were looking at and instead merely classified and labelled the target's expression, perhaps on the (correct) assumption that this cognitively simpler strategy would yield an equally accurate response.

The hypothesis derived from the theory of retrodiction is as follows: When asked which expression the target is viewing, perceivers first look for signals in the target's behavior, enabling them to classify the target's facial expression if possible, and then sense the mental reaction of the target (i.e. their inner state) followed by an inference of the expression being viewed by the target. The opposing hypothesis says that no inference is involved and that processing ceases upon the perceiver classifying the target's expression. This hypothesis was duly tested by Kang et al (2018) in which perceivers wore EEG nets under two experimental conditions, one in which perceivers were asked which expression the target was viewing (the retrodiction question) and one in which perceivers were asked to classify the target's expression (without reference to anything the targets might be viewing). The retrodiction hypothesis predicts that when perceivers answer the retrodiction question the cortical signal will sustain for a longer period (reflecting the additional mental operation involved) and will reveal activity in areas reputed to be involved in mindreading, compared with when they answer the classification question. The opposing hypothesis predicts identical cortical signatures (in duration and in location) when perceivers answer both questions.

In support of the hypothesis, Kang et al (2018) found that when perceivers were answering the classification question this involved little more than processing the target's expression, marked by activity in the fusiform face area (Kanwisher & Yovel, 2006). In contrast, answering the retrodiction question gave rise to additional cortical

activity associated with processing inner states, particularly in the medial prefrontal cortex and in the left and right temporo-parietal junction (Schurz et al., 2014). Compared with answering the classification question, and considering the findings of Sabbagh et al (2004), EEG measurements revealed stronger slow-wave activity (N270-400) over the inferior frontal and anterior temporal regions of the right hemisphere. Hence, the evidence suggests retrodiction amounts to more than classification of a target's facial expression.

**5. Individual differences in readability**

Past research into retrodiction has mostly focused on how accurately *perceivers* make judgments about targets, especially the stimuli or events they are reacting to. But because multiple perceivers make such judgments about each target, there is a mean score associated with each target representing the average level of accuracy achieved by a group of perceivers in relation to judgments of what caused that particular target's reaction. This could be defined as a readability value for that target. Evidently, the prospects of accurate mindreading depend not only on the ability of the perceiver to carry out the appropriate processing but also on the quality of signal emanating from the target. The notion that some targets may have inner states that are more readable mirrors ideas in the Realistic Accuracy Model (Funder, 2012) which proposes that there are individual differences in target readability in relation to personality judgments.

The procedure described above, that stratifies the three principles, offers an additional benefit in the sense that it lends itself to coding at the level of the target, to determine individual differences in target readability. We can thus place targets in rank order from most to least readable, with the expectation that some will be significantly readable, some will be unreadable and some will emit signals that are significantly misleading. Unreadable targets are those for whom perceivers are unable to make accurate inferences without there being any consensus on an incorrect inference. Misleading targets are not only unreadable but are systematically misleading to perceivers, who are liable to infer that targets are reacting to something other than they are actually reacting to. For example, in our pilot work, a particular target was videoed while reacting to the cue word 'anger' and yet perceivers tended to infer that he was viewing the cue word, 'happy' apparently due to his smiling expression (albeit, perhaps a devilish smile!).

Recent research shows that autistic targets tend to be very difficult to read by neurotypical perceivers (Alkhaldi et al, 2019; Sheppard et al, 2016). And targets who vary on the trait empathy continuum vary in how readable they are, depending upon what kind of thing they are thinking about (Wu & Mitchell, 2019). Collectively, these studies suggest an expressive target is not necessarily readable; the quality rather than the strength of the signal emanating from the target is critical. Hence, the link between readability and personality is complicated by the particular inner state the target is experiencing.

## 6. Summary and conclusion

In this article we propose a framework for investigating how accurately and by what process people read others' minds – a process that requires perceivers to make a retrodictive inference. In other words, after witnessing a target's behavior, we assume the perceiver is sensitive to a proximal cause of target behaviour (the target's inner state) which allows them to proceed to a distal inference (the event in the world that caused the target's inner state). The perceiver is asked to infer the causal event in the world and if they can do this then we assume they have also sensed the target's inner state.

This approach complies with the three principles outlined above. First, the behavior of the target (and their signaled inner state) is spontaneous (Principle 1): targets are not acting according to instruction but are spontaneously reacting, sometimes quite subtly, to a worldly event. Second, the truth condition is satisfied in that the perceiver's judgment (and by implication their accuracy in mindreading) can be compared against an objective fact (Principle 2). Third, the judgments made by perceivers lend themselves to coding using either signal detection or Wagner's unbiased hit rate, codings which quantify response bias and which duly separate accuracy from bias (Principle 3). This approach embraces the idea that mindreading is adaptive by, amongst other useful things, enabling people to know the world through the lens of other minds.

References

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

Al Khaldi, R., Sheppard, E. & Mitchell, P. (2019). Is there a link between autistic people being perceived unfavourably and having a mind that is difficult to read? *Journal of Autism and Developmental Disorders*, *49*, 3973–3982. doi:10.1007/s10803-019-04101-1

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

Archer, D., & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology*, *35*, 443-449. doi:10.1037/0022-3514.35.6.443

Back, E., & Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, *115*, 54–70. doi: 10.1016/j.cognition.2009.11.008

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

Bartsch, K., & Wellman, H. (1989). Young children's attribution of action to beliefs and desires. *Child Development*, *60*, 946-964. doi:10.2307/1131035

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382-386. doi:10.1111/j.1467-9280.2007.01909.x

Birdwhistell, R. (1952). Introduction to kinesics: (An annotation system for analysis of body motion and gesture). Louisville, KY: University of Louisville.

Birdwhistell, R. (1970). Kinesics and context. Philadelphia: University of Pennsylvania Press.

Cassidy, S., Ropar, D., Mitchell, P. & Chapman, P. (2013). Can adults with autism spectrum disorders infer what happened to someone from their emotional response? *Autism Research*, *7*, 112-123. doi:10.1002/aur.1351

Cassidy, S., Ropar, D., Mitchell, P., & Chapman, P. (2015). Processing of spontaneous emotional responses in adolescents and adults with Autism Spectrum Disorders: effect of stimulus type. *Autism Research*, *8*, 534-544. doi:10.1002/aur.1468

Dziobek, I, Fleck, S., Kalbe, E., Rogers K., Hassenstab, J., et al. (2006). Introducing

MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*, 623-636. doi:10.1007/s10803-006-0107-0

Epley, N., Morewedge, C., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*, 760-768. doi:10.1016/j.jesp.2004.02.002

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: *Human Perception and Performance*, *1*, 288-299. doi:10.1037/0096-1523.1.3.288

Flavell, J. H., Miller, P. H., & Miller, S. A. (1993). Cognitive development (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Fodor, J. A. (1992). A theory of child's theory of mind. *Cognition*, *44*, 283-296. doi:10.1016/0010-0277

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*, 177-182. doi:10.1177/0963721412445309

Gallagher, S. (2011). The overextended mind. *Versus: Quaderni di studi semiotici*. *113*, 57-68.

Gallese, V. (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, *4*, 23–48. doi:10.1007/s11097-005-4737-z

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Cognition*, *2*, 493-501. doi:10.1016/S1364-6613(98)01262-5

German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in "theory of mind": Evidence from compromised belief-desire reasoning in old age. *Cognition*, *101*, 129-152. doi:10.1016/j.cognition.2005.05.007

Goldman A, & de Vignemont, F (2009). Is social cognition embodied? *Trends in Cognitive Sciences*, *13*, 154–159. doi:10.1016/j.tics.2009.01.007

Grynberg, D., Chang, B., Corneille, O., Maurage, P., Vermeulen, N., Berthoz, S. et al. (2012). Alexithymia and the processing of emotional facial expressions (EFEs): Systematic review, unanswered questions and further perspectives. *PLoS ONE*, *7: 8*, e42429. doi:10.1371/journal.pone.0042429

Heavey, L., Phillips, W., Baron-Cohen, S., & Rutter, M. (2000). The awkward moments test: A naturalistic measure of social understanding in autism. Journal of *Autism and Developmental Disorders*, *30*, 225-236. doi:10.1023/A:1005544518785

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*, 131-143. doi:10.1177/1745691613518076

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*, 730-742. doi:10.1037/0022-3514.59.4.730

Ickes, W. (2003). Everyday mind reading: Understanding what other people think and feel, Amherst, NY: Prometheus Books.

Ickes, W. (2009). "Empathic accuracy: its links to clinical, cognitive, developmental, social, and physiological psychology," in *The social neuroscience of empathy*, eds. J. Decety & W. Ickes (Cambridge, MA: MIT Press), 57-70.

Johnson, C. N., & Maratsos, M. P. (1977). Early Comprehension of Mental Verbs: Think and Know. *Child Development*, *48*, 1743-1747. doi:10.2307/1128549

Johnson, C. N., & Wellman, H. M. (1980). Children's developing understanding of mental verbs: Remember, know and guess. *Child Development*, *51*, 1095-1102. doi:10.2307/1129549

Kang K., Anthoney L. & Mitchell P. (2017). Seven- to 11-year-olds' developing ability to recognize natural facial expressions of basic emotions. *Perception*, *46*, 1077-1089. doi:10.1177/0301006617709674

Kang, K., Schneider, D., Schweinberger, S., & Mitchell, P. (2018). Dissociating neural signatures of mental state retrodiction and classification based on facial expressions. *Social Cognitive and Affective Neuroscience*, *13*, 933–943. doi:10.1093/scan/nsy061

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. Philosophical Transactions: *Biological*

*Sciences*, *361*, 2109-2128. doi:10.1098/rstb.2006.1934

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Sciences*, *11*, 32-38. doi:10.1111/1467-9280.00211

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

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

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Mitchell, P., Robinson, E. J., Isaacs, J., & Nye, R. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition*, *59*, 1-21. doi:10.1016/0010-0277(95)00683-4

North, M. S., Todorov, A., & Osherson, D. N. (2010). Inferring the preferences of others from spontaneous, low-emotional facial expressions. *Journal of Experimental Social Psychology*, *46*, 1109-1113. doi:10.1016/j.jesp.2010.05.021

North, M. S., Todorov, A., & Osherson, D. N. (2012). Accuracy of inferring self- and other-preferences from spontaneous facial expressions. *Journal of Nonverbal Behavior*, *36*, 227-233. doi:10.1007/s10919-012-0137-6

Pillai, D., Sheppard, E., Mitchell, P. (2012). Can people guess what happened to others from their reactions? *PLoS ONE*, *7: 11*, e49859. doi:10.1371/journal.pone.0049859

Pillai, D., Sheppard, E., Ropar, D., Marsh, L., Pearson, A., & Mitchell, P. (2014). Using other minds as a window onto the world: Guessing what happened from clues in behaviour. *Journal of Autism and Development Disorders*, *44*, 2430-2439. doi:10.1007/s10803- 014-2106-x

Robinson, E. J., & Mitchell, P. (1995). Masking of children's early understanding of the representational mind: backwards explanation versus prediction. *Child Development*, *66*, 1022-1039. doi:10.2307/1131796

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology*, *9*, 331-349. doi:10.1111/j.2044-835X.1991.tb00881.x

Sabbagh, M. A., Moulson, M., Harkness, K. (2004). Neural Correlates of Mental State Decoding in Human Adults: An Event-related Potential Study. *Journal of Cognitive Neuroscienc*. *16*, 415-426. doi:10.1162/089892904322926755

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9-34. doi:10.1016/j.neubiorev.2014.01.009

Sheppard, E., Pillai, D., Wong, G. T., Ropar, D., & Mitchell, P. (2016). How easy is it to read the minds of people with Autism Spectrum Disorder? *Journal of Autism and Developmental Disorders*, *46*, 1247-1254. doi:10.1007/s10803-015-2662-8

Sorce, J. F., Emde, R. N., Campos, J. J., & Klinnert, M. D. (1985). Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology*, *21*, 195-200. doi:10.1037/0012-1649.21.1.195

Teoh, Y., Wallis, E., Stephen, I.D. & Mitchell, P. (2017). Seeing the world through other minds: Inferring social context from behaviour. *Cognition*, *159*, 48-60. doi:10.1016/j.cognition.2016.11.003

Valanides, C., Sheppard, E., & Mitchell, P. (2017). How accurately can ohter people infer your thoughts—And does culture matter? *PLoS ONE*, *12(11)*: e0187586. doi:10.1371/journal.pone.0187586

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*, 3-28. doi:10.1007/BF00987006

West, T. V. & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*, 357-378. doi:10.1037/a0022936

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

Wu, W., Sheppard, E., & Mitchell, P. (2016). Being Sherlock Holmes: Can we sense empathy from a brief sample of behaviour? *British Journal of Psychology*, *107*, 1-22. doi:10.1111/bjop.12157

Wu, W., Sheppard, E., & Mitchell, P. (2017) Judging personality from a brief sample of behavior: It is relatively easy to detect who is unique on a trait. *European Journal of Personality*, *31*, 685-700. doi:10.1002/per.2116

Wu, W., & Mitchell, P. (2019). Accurate inferences of others' thoughts depend on where they stand on the empathic trait continuum. *Personality and Individual Differences*, *148*, 110-116. doi:10.1016/j.paid.2019.05.025

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science*, *19*, 399-404. doi:10.1111/j.1467-9280.2008.02099.x

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion*, *9*, 478-487. doi:10.1037/a0016551

Zaki, J. & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry*, *22*, 159-182. doi:10.1080/1047840X.2011.551743

Footnotes

1. Hypothetically, suppose the target declared something that is not their true mental state; how, then, could we explain cases where the perceiver estimates the target holds such a mental state that actually corresponds with the target's declaration? In this kind of task, the target watches a video of their own behavior and rates their mental state at any given time while watching. If their mind were not transparent to itself, then presumably they would effectively view themselves as if viewing a third party and infer mental states from clues in (their own) observable behavior. This is precisely what perceivers are asked to do and therefore it is hardly surprising to find that the kind of judgments perceivers make on observing the target coincide with the declarations made by the target. Hence, perceiver success in this task does not qualify as evidence that the truth condition has been satisfied.

2. The trouble, though, is we are left with the lingering doubt that when the perceiver retrodictively inferred that the target's reaction was caused by Event X, they made a direct link between target behavior and the external causal event without giving any consideration to the target's inner states. We will return to this problem subsequently and suggest that the experimental procedure lends itself to using the tools of cognitive neuroscience to illuminate whether or not perceivers are indeed sensing targets' inner states.

Figure legends

Figure 1. Three different models of how perceivers might infer the cause (an event in the world) of a target's reaction. Figure 1a depicts a causal chain in which the perceiver sesnes the target's mental state from their reaction (Step 1) and then infers the event that caused the mental state (Step 2). Figure 1b depicts the perceiver sensing the target mental state and the event in the world that caused the target's reaction, but inferring one (the event in the world) is not dependent on the other (sensing the mental state). Figure 1c depicts the perceiver inferring the event in the world that caused the target's reaction without sensing the target's mental state at all.

# Figures

Figure 1a

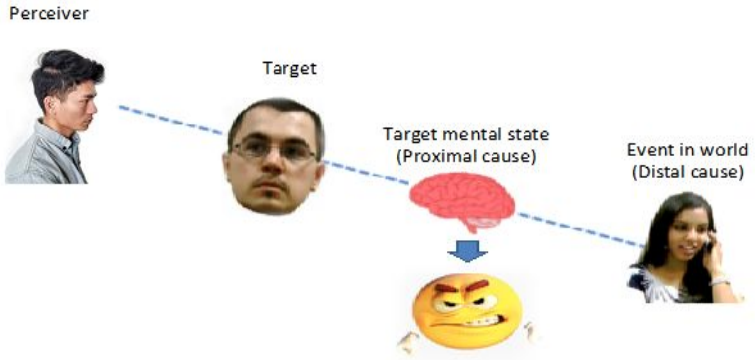


Figure 1b

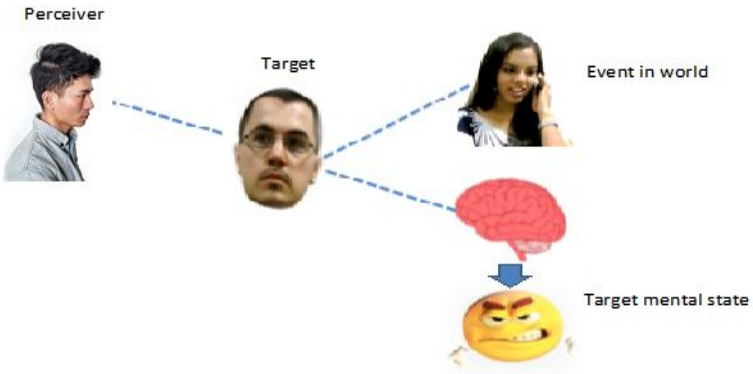


Figure 1c