Erkenntnis
https://doi.org/10.1007/s10670-018-0090-9

ORIGINAL RESEARCH

Prediction Error Minimization as a Framework for Social Cognition Research

Leon de Bruin[1] · **John Michael**[2]

Received: 10 July 2017 / Accepted: 19 November 2018

Abstract
The main aim of this article is to give an assessment of prediction error minimization (PEM) as a unifying theoretical framework for the study of social cognition. We show how this framework can be used to synthesize and systematically relate existing data from social cognition research, and explain how it introduces new constraints for further research. We discuss PEM in relation to other theoretical frameworks of social cognition, and identify the main challenges that this approach to social cognition will need to address.

1 Introduction

How do we understand others? For a long time, research on social cognition was dominated by two positions. According to the first position, understanding others requires a folk psychological theory—a 'theory of mind'—that allows us to explain and predict their behavior. This position is also known as the Theory Theory (TT). The second position, by contrast, claims that we understand others by 'putting ourselves in their shoes' and by simulating the mental states we would have in their situation. This is the Simulation Theory (ST). Both positions were quite successful in generating testable hypotheses and structuring research in developmental psychology, moral psychology, social neuroscience and comparative psychology.

After several decades of intense debate, however, the consensus seems to be that neither TT nor ST can provide a unifying theoretical framework for the study of social cognition (Apperly 2008, 2011). As a result, many researchers in the field have embraced an alternative framework, such as a hybrid of TT/ST (e.g., Nichols and Stich 2003; Goldman 2006), a dual-systems theory (e.g., Apperly and Butterfill

✉ Leon de Bruin
l.debruin@ftr.ru.nl

[1] Department of Philosophy, Radboud University Nijmegen, Erasmusplein 1, Postbus 9103, 6500 HD Nijmegen, The Netherlands

[2] Department of Philosophy, University of Warwick, Coventry CV4 7AL, UK

Published online: 30 November 2018

2009; De Bruin and Newen 2012; Fiebich 2014), or a multi-systems framework (Christensen and Michael 2013).

However, new developments in cognitive neuroscience and computational modeling, in particular the emergence of the Prediction Error Minimization (PEM) framework, suggest that this move might be premature. One of the key selling points of PEM is its ability to integrate a broad range of findings into a single coherent framework. The main aim of this article is to assess whether PEM can provide a unifying theoretical framework for the study of social cognition.

The article is structured as follows. In Sect. 2 we introduce the basic principles and assumptions that underlie PEM. Sections 3 and 4 focus on the application of the PEM framework to research on social cognition. We show how PEM sheds new light on existing findings, in particular those on false belief understanding, and how it can be used to raise interesting new questions for further research. In Sects. 5 and 6, we discuss PEM in relation to other theoretical frameworks. Finally, in Sect. 7, we conclude by identifying the main challenges that a PEM approach to social cognition will need to address.

2 Prediction Error Minimization: A Primer

Before discussing the basic principles and assumptions of PEM, we want to point out that although PEM is a formal apparatus that can be used to generate precise quantitative models, it is also a loose general framework in the sense that it can be used in the service of different explanatory programs.[1] It is only by enriching this loose general framework with additional assumptions about a specific domain (such as visual perception, motor control or social cognition) that precise quantitative models, and thus also precise testable predictions, can be generated. Our strategy will be to begin with a brief characterization of the general framework, and then to consider how to enrich it with additional assumptions specific to the domain of social cognition (Sects. 3, 4).

PEM conceives of the brain as a probabilistic inference system, which attempts to predict the input it receives by constructing models of the possible causes of this input. The main aim of this system is to minimize the 'prediction error'—the discrepancy between the predicted and the actual input. If the prediction error is small, then there may be no need to revise the model that gives rise to the prediction. If, on the other hand, the prediction error is large, then it is likely that the model fails to capture the causes of the inputs, and therefore must be revised. In this sense, the brain is not concerned with coding input per se but only *unexpected* input. This is nicely illustrated in the area of reward processing by the behavior of dopaminergic neurons in the striatum: their rate of firing corresponds to unexpected changes in the value of a coming reward (e.g. increases or decreases in the number of drops of juice

[1] The contrast between the formulations of PEM found in Hohwy (2013) and Clark (2013, 2016) serves to illustrate this theoretical openness of PEM.

that are administered after a tone has sounded), not to the actual value of the reward itself (Bayer and Glimcher 2005; Nakahara et al. 2004; Tobler et al. 2005).

PEM postulates that the models generated by the brain are not only evaluated according to how well they fit the evidence, i.e., how well they predict the input in question, but also according to how likely they are in the first place, i.e., their 'prior probability'. When making sense of new input the brain does not start from scratch, but rather updates the models with the highest prior probability in order to accommodate the new evidence.[2]

Furthermore, PEM assumes that these models are organized in a *hierarchy*. At the lowest level of the hierarchy, neural populations encode such features as surfaces, edges and colors. At a hierarchically superordinate level, these low-level features are grouped together into objects, while even further up the hierarchy these objects are grouped together as components of larger scenes involving multiple objects. When you see a red teacup, for example, there will be a response on the part of neurons in your visual system that code for edges, and these neurons will represent edges at a particular location in the visual field. In addition, there will be a response on the part of neurons that code for surfaces, and there will be a response on the part of neurons that code for redness, which will represent a surface and redness at a particular area of the visual field. From one millisecond to the next, there will not be much change in these inputs, and the neural populations at the hierarchically lowest level (representing edges, surfaces and colors) may, as a default, predict no change in inputs. If the cup is moved, however, the inputs will change. Importantly, they will change in a manner that is coherent, given that they are all features of the same cup—if one of the edges moves to the left a certain distance, so will the other edges, and so will the red surface. In order to draw upon such regularities in anticipating inputs, the brain, at a hierarchical level that is superordinate to the representation of such low-level features as surfaces and edges and colors, represents the cup as an object. Moreover, to anticipate changes over longer time scales, superordinate models embed this object into larger scenes, such as tea parties, and thereby generate predictions pertaining to objects and overall scenes in a context-dependent fashion (rather than low-level features such as edges, surfaces and colors). Thus, by embedding the cup into a model of a tea party, it will become possible to predict roughly in what ways the cup will be moved, by whom, and where to. On the other hand, since we also lose detail and precision as we move up the hierarchy, lower hierarchical levels are still required in order to make specific predictions.

Finally, PEM proposes that the brain has two options for reducing prediction error. The first option is to revise its model of the world until the prediction error is satisfactorily diminished ('perceptual inference'). The second option is to change the world so that it matches the model ('active inference'). If, for example, one expects to see one's teacup on the desk in front of one, but it turns out not to be there, one

[2] In the Bayesian interpretation of predictive coding estimating the causes comes down to finding the most probable causes vm given the input u for that level and the current model parameters θ (Friston 2002; Blokpoel et al. 2012):
$vm = \arg\max_v \Pr(v|u;\theta)$.

might simply conclude that one was mistaken (i.e. change the model). But one might also adjust one's head or even one's bodily position until one does see the cup, e.g. behind the laptop or occluded by a stack of books. In this case, one has changed the world in the sense of changing the position of one's body in the world. More radically, one might *go and get a cup of tea* and put it exactly on that part of the desk where one had expected it to be. Again, this would amount to changing the world to match the model one had of it (Friston et al. 2011).

Having laid out the basic principles and assumptions behind the PEM framework, let us now consider how they might be applied in the study of social cognition.

3 Applying PEM to Social Cognition Research

In considering how to apply the PEM framework to social cognition, one might start out from the observation that other people are a particular kind of cause of sensory input on a par with teacups. Just as it is sometimes helpful in reducing uncertainty about the behavior of edges and surfaces and colors to postulate teacups as the bearers of those lower-level features, it is also sometimes useful to postulate other people. However, the differences between people and cups are at least as interesting as the similarities, and this is what motivates the project of examining social cognition as a particular subset of cognition in general.

Various differences already manifest themselves at the lowest levels of the hierarchy, as there are neural populations in early visual areas that specialize in detecting specific features of agents, such as faces (Haxby et al. 2002), eyes (Haxby et al. 2000), gaze direction (Teufel et al. 2009) and emotions (Tamietto et al. 2009). These features trigger the brain to bring a whole set of expectations to bear that are specific to agents in contrast to non-agentive objects. To focus for a moment just on the example of eyes, there is a very deep-seated expectation that one will learn something useful by following others' gaze direction. Gaze following occurs by 6 months at the latest (Senju and Csibra 2008), and Hood et al. (1998) have even found evidence for it in 2.5 month-olds. And if another agent's eyes happen to be directed at oneself, then (s)he will be able to acquire information about one's emotions from one's facial expressions and posture, to track the direction of one's gaze and to anticipate one's movements. Moreover, it is likely that (s)he will initiate certain actions directed towards one. Thus, it should come as no surprise that eye contact consistently elicits a relatively strong reaction, as evinced, for example, by galvanic skin response, electroencephalogram activity (Nichols and Champness 1971; Gale et al. 1975), and the activation of specific motivational brain systems (Hietanen et al. 2008). In fact, the importance of eye contact is so fundamental and pervasive that even newborns are sensitive to it. Farroni and colleagues, for example, found that 2- to 5-day-old newborns looked longer and more frequently at a photograph of a face whose eyes were facing directly to them than a different image of the same face with the eyes facing away (Farroni et al. 2002).

It is interesting in this connection to consider the results of Samson et al. (2010)'s well-known study on automatic level-1 perspective taking. In this experiment, the perspective of an avatar interferes with the ability of participants to judge how many

objects they (the participants) could see, suggesting that the participants calculate how many objects the avatar can see (Samson et al. 2010; see also Qureshi et al. 2010). The authors argue that this effect is driven by an automatic perspective-taking process. A different possibility, however, is that the gaze direction of the avatar functions as a cue, which directs attention to and facilitates the processing of objects at the cued location. The latter interpretation gains support from a recent study by Santiesteban et al. (2013), who replicated the effect with arrows rather than avatars. This suggests that the effect may not be driven by (automatic) perspective taking processes but by attentional cueing. On the basis of the PEM framework, one might speculate that the avatar and the arrows have this effect because there is a prior probability that if a location is cued by an arrow, or if an agent is looking in one direction, it is likely to be worthwhile to look there (i.e., we would expect high precision prediction error there). If so, then it should be possible to modulate the effect by changing the prior probability that the avatar's gaze direction will be a useful cue to the location of the targets. For instance, the interference effect should increase if the avatar is believed to be highly competent, and it should decrease if she or he is believed to have poor vision. Thus, whereas automatic perspective-taking and attentional cueing views predict that background knowledge should make no difference (since the relevant process is entirely stimulus driven, either by domain specific representations of perceptual perspectives or domain general directional cues), PEM predicts that the right kind of background knowledge should moderate the interference effects.

The above discussed findings, seen through the lens of the PEM framework, suggest that the very appearance of a human-like body with features such as eyes triggers a cascade of predictions. However, one might wonder how these processes get off the ground in the first place. For example, it might be objected that the developmental findings by Farroni et al. (2002) actually pose a problem for the PEM framework, because it is not immediately obvious that the neonatal preference for direct eye contact is based on a prediction or driven by priors. Thus, one might argue that attentional biases for face-like shapes in early ontogeny are in fact automatic, stimulus-driven responses to cues in the environment—the product of innate, domain-specific mechanisms for social learning. But PEM is not necessarily incompatible with such an account. For example, it could be argued that neonatal attentional biases have a scaffolding function in that they allow children to inductively infer generalizations about the relation between eye-movements and what happens in the surrounding environment. These generalizations can then support prediction error mimization processes in later development.[3] On its own PEM does not actually help us to decide whether the effects in the dot perspective task are driven by domain specific representations of perceptual perspectives or domain general directional cues. This illustrates our point that PEM is a general framework that needs to be enriched with additional assumptions about a specific domain to generate interesting hypotheses.

[3] We thank an anonymous reviewer for this suggestion.

Human bodies not only look different from teacups, they are also subject to characteristic regularities governing their movements that do not apply to objects in general—specifically, human bodies are subjects to biomechanical laws. Thus, by hypothesizing the presence of a human body, the brain can bring a whole new set of expectations to bear in predicting upcoming sensory inputs. There is evidence that some neural populations, especially in the superior temporal sulcus, respond to biological motion. Moreover, the (perceptions of) movements that violate these laws (such as a finger bending too far backward) induce a greater response in the superior temporal sulcus than biomechanically predictable movements (Costantini et al. 2005). Similarly, the perception of a human-like body which, surprisingly, moves about in an inhuman, mechanical manner elicits a greater response (prediction error signal) than a human-like body moving about in a human-like fashion or a non-human-like body moving about in a non-human-like manner (Saygin et al. 2012). The brain also utilizes a set of expectations about how humans (and other animals) move through space. If a person is walking across the room and passes behind a barrier, the default prediction is that they will continue along the same path and re-appear on the other side of the barrier moving at the same speed. Again, it is the superior temporal sulcus that appears to specialize in generating predictions at this level (Saxe et al. 2004).

Moving up to a hierarchically superordinate level, the brain embeds human bodies into models of larger situations, thereby making it possible to draw upon knowledge about what behavior is likely in light of contextual factors. For example, a prediction with high prior probability is that people turn toward objects that suddenly appear or suddenly emit bright lights (Pelphrey et al. 2003; Pelphrey and Van der Wyk 2011). Moreover, by representing others people as actors within situations, it is possible to appeal to situational regularities that derive from the fact that people's behavior, unlike that of coffee cups and clouds, is governed by various kinds of norms and conventions. Social and moral norms, for example, make it highly unlikely that people will run up and kiss strangers, strangle cats in public squares, or drive their cars on the left side of the road (unless they happen to be in one of the countries in which conventions dictate that they do so). Sensitivity towards these norms already emerges in early ontogeny. Developmental studies show that when infants begin to imitate the actions of adults, they do not merely re-enact the idiosyncratic act of individuals, but rather learn something about the general form of the action, and how normative dimensions of appropriate and inappropriate performance structure it. For example, when 2-year-old infants see someone use a novel object systematically in an instrumental way, they use the object in similar ways themselves later on, and only for this purpose. What is more, they expect other people to do so as well (Casler and Kelemen 2005). In a series of studies, Rakoczy and colleagues demonstrated that, in addition to respecting social norms, infants also enforce them on third parties. Not only do they protest at norm violations, but they also try to alter the norm transgressor's behavior, for instance, by teaching the 'right' way to do it

(Rakoczy 2008; Rakoczy et al. 2008). This could very well be construed as a form of active inference: children minimize prediction error by correcting behavior that is 'out of line' and by bringing it in line with their expectations about how agents should behave in specific situations.[4] Furthermore, the results also seem to indicate that children acquire the higher-order expectation that behavior conforms to social norms and use this to infer particular norms in their social environment. This could be interpreted as a case of *overhypothesis learning*, which focuses on the constraints on the hypotheses considered by the learner. Kemp et al. (2007) have shown how hierarchical Bayesian models can help to explain how overhypotheses about feature variability (e.g. the shape bias in word learning) and the grouping of categories into ontological kinds (e.g., objects and substances) are acquired during development. Perhaps a similar kind of explanation might be given for overhypotheses about behavior variability in a social context.[5]

Moving still further up the hierarchy, there are also psychological features that are unique to specific agents and which persist over longer time scales, such as their preferences. Thus, agents can be expected to act in ways that are consistent with their specific preferences. For example, given that a person who smiles at a mug and frowns at a stuffed animal is more likely to reach for the mug than the stuffed animal, it will be surprising to observe the person to opt for the stuffed animal. This is inconsistent with their inferred preference, and indeed the superior temporal sulcus is responsive to this surprising event (Van der Wyk et al. 2009).

Similarly, the ascription of specific beliefs and desires to individual agents makes it possible to exploit regularities that can span quite long timescales. If an agent believes that the object she desires is at one location, she is likely to seek it there regardless of whether it really is still there, and indeed even if it happens to have been transferred years ago to some other location. Recent findings in developmental psychology suggest very strongly that infants are able to anticipate such regularities by the second year of life (Baillargeon et al. 2010; Butterfill and Apperly 2013), and perhaps even by 6–7 months (Kovács et al. 2010; Southgate and Vernetti 2014). And when, in experimental settings in which false belief scenarios are generated, this type of pattern is broken, children tend to look longer at the scene, suggesting that their expectation has been violated.[6]

Much has been made of the discrepancy between these findings, on the one hand, and the failure of 3–4 year-old children to succeed at false belief tests in which they are asked verbally to predict where an agent with a false belief is likely to search for the object in question (Wimmer and Perner 1983; Wellman et al. 2001; Butterfill and Apperly 2013). Attempts to account for this developmental puzzle have ranged from denial that the data really does reveal a sensitivity to belief states on the part

[4] Although we will sometimes offer interpretations of behavior that are couched in agential terms (e.g., children who "minimize prediction error" and "engage in active inference"), we assume that PEM processes take place at the sub-personal level.

[5] We thank an anonymous reviewer for this suggestion.

[6] Many other measures apart from looking time have also been used. For reviews, see Baillargeon et al. (2010) and Christensen and Michael (2013).

of infants (Heyes 2014; Perner and Ruffman 2005), to arguments that the infants do represent beliefs per se but have specific difficulties with various task demands arising in explicit verbal versions (Baillargeon et al. 2010) to something in between (Apperly and Butterfill 2009; De Bruin and Newen 2012).

In what follows, we will highlight some points that are raised by the PEM framework and which may contribute to a satisfying theoretical account of these findings.

4 PEM and False Belief Understanding

A first point worth mentioning is that the PEM framework is neutral with respect to the question whether infants have to *meta-represent*, i.e. to represent another agent's belief about the world (specifically, the location of an object) in order pass these tests. The results of the violation-of-expectation false belief test, for example, indicate that infants form expectations about the behavior of another agent, which are violated at some point during the experiment. Similarly, the results of the anticipatory-looking false belief test suggest that infants form certain expectations about the behavior of another agent, which are manifested in their looking behavior. From a PEM perspective, these findings must be explained by specifying how exactly the relevant expectations are generated. For example, as was mentioned in the previous section, it might be the case that these expectations are the product of innate, domain-specific mechanisms, which allow infants to infer generalizations about the relation between eye-movements and what happens in the surrounding environment. But PEM itself does not commit us to assuming that representations of the agent's beliefs provide the best explanation; it could for example be that something like registrations in Butterfill and Apperly's (2015) sense provide a better explanation, or a radical ecological-enactive approach (Bruineberg et al. 2016). In any case, PEM does not resolve this dispute so much as provide a different set of terms and principles in which to think about it. However, this does not mean that PEM cannot offer novels insight on some aspects of the infant mindreading puzzle.

One fresh starting point which PEM offers focuses on the question whether and how infants deal with the prediction error that follows when their expectations are violated. With this question in mind, it is worth taking a closer look at experiments using *active-helping* as a dependent measure. These studies show that children are not merely surprised when their expectations about another agent's behavior are violated but actively try to reduce the resulting prediction error by assisting the agent in question.[7] That is, children engage in *active inference* by changing the world so that it matches their model of it (Michael et al. 2014; Cf. Michael and Székely 2017). For example, the experiment by Buttelmann et al. (2009) shows that 18-month-olds, when confronted with a mismatch between (their representation of) the preference of the agent and the new location of the object, reduce this prediction error by directing the agent to the actual location of the object and extracting the object for him.

[7] Note that the term 'surprise' as we use it here refers to *surprisal* (the implausibility of some sensory state given a model of the world) rather than *agent-level surprise.*

Young children's superior performance in experiments based on active helping compared to elicited response paradigms may provide an important clue for explaining the developmental puzzle of false belief understanding. Most accounts of false belief understanding (e.g., Baillargeon et al. 2010; Carruthers 2016) assume that children have to inhibit their own true belief and select and represent the false belief of the other agent in order to correctly predict the action that follows. These accounts hypothesize that young children fail the elicited-response false belief test because they cannot handle the complexity of the task and therefore default to what they themselves believe to be the case. PEM, however, suggests that there might be a different reason why younger children in the elicited-response false belief test *systematically give incorrect answers*, rather than confused or arbitrary answers. Specifically, it suggests the idea that the elicited-response false belief test is more difficult because it confronts children with a prediction error that cannot be reduced by perceptual inference, i.e. by updating their model of the agent (specifically the agent's belief about the location of the object). In the active-helping false belief test, children can reduce the prediction error by means of active inference. Indeed, this is the most effective way to reduce the prediction error. In the elicited-response false belief test, by contrast, this option is not available. If we assume that children have a natural inclination to reduce prediction error, then the elicited-response false belief test might be more difficult because young children need to inhibit their tendency to engage in perceptual inference and update their model of the agent's belief to get rid of the discrepancy between the preference of the agent and the new location of the object. If correct, this explanation would provide a partial reduction of the explanation recently offered by Helming et al. (2014): they suggest that, since young children want to help others, they make the pragmatic error of trying to be informative in order to help in elicited-response tasks. The current alternative explanation would suggest that in fact kids may want to help in order to reduce prediction error—and thus that they may also make the error in a competitive setting (whereas Helming and colleagues' hypothesis would not predict this). In other words, they may often help others non-strategically, e.g. in game situations in which it would not be in their interest to do so.

Obviously, our proposal would require further development. In particular, some account must be given of *when* infants help. After all, infants and young children do not *always* help with others' goals. Indeed, they show preferences in whom they help, preferring not to help agents with harmful intentions (Vaish et al. 2010; Dunfield and Kuhlmeier 2010; Dahl et al. 2013). To some extent, it may be possible to explain these findings in terms of other factors which are compatible with the active inference account—e.g. there is no reason why the active inference account should not incorporate a preference to interact with agents who have shown kindness, or with specific people, ingroup members, etc. Indeed, infants are probably more likely to interact with an agent who has shown kindness than with an agent who has not. If so, then active inference would be a probable means of reducing prediction error

in the former case but not in the latter. And since prediction errors, on this account, will always be reduced by whatever means is most probable, active inference (i.e. active helping) may be the most probable means in cases in which the agent has shown kindness and/or is familiar, but not in other cases.[8]

5 Situating PEM in Contemporary Debates on Social Cognition: Mindreading

In order to assess the merits of PEM as a research framework, it is important to determine whether it can address (some of) the central issues with the existing accounts of social cognition. Our aim in this section is not to provide a complete overview of these issues or to explain them in detail. Instead, we will simply focus on those issues that PEM might be able to illuminate to some extent.

One of the main controversies in the contemporary debate on social cognition concerns the centrality of mindreading. Various arguments have been put forward to show why overestimating the relative importance of mindreading for social cognition is problematic. For example, Bermúdez (2003) has claimed that Theory Theory (TT) and Simulation Theory (ST) accounts of mindreading are unattractive because they are highly demanding in terms of computational complexity. TT is computationally complex because it requires agents to figure out which theoretical principles could apply in a given situation, whether appropriate background conditions hold, whether there are countervailing factors, etc. ST is computationally complex because agents have to identify the mental states that could be relevant to the simulation process, run pretend versions of these states through their own decision-making mechanisms, and attribute the outcome (the pretend decision) to the target. According to Bermudez, this is problematic because most of our social interactions are relatively 'smooth' and involve "almost instantaneous adjustments to the behavior of others" (2003, p. 8).

A similar objection might be leveled at proponents of a PEM account of social cognition. As far as we can see, there are basically two ways in which they could respond. On the one hand, they could argue that even though socio-cognitive processes are computationally complex, this does not mean that we have to *experience* them as such. That is, following Clark, they could argue that there is "clearly no inconsistency in thinking that the brain's pervasive use of probabilistic encoding might yield conscious experiences that depict a single, unified and quite unambiguous scene." (2013, p. 16; see Spaulding 2010 for a similar defense of TT and ST)

On the other hand, however, proponents of PEM could also point out that inferences about other agents' mental states do not need to be *concurrently* involved in the prediction of their behavior—as long as the priors are adequate. This is because the brain does not need to engage in rich inferential processing as long as

[8] It must be acknowledged, of course, that other accounts would also generate this same prediction. In order to test the active-inference account of active-helping, it would be necessary to identify predictions which it uniquely generates. This is an important objective for future research.

expectations about another agent's actions are (more or less) met. Instead, it is only *unexpected* input that generates error signals and thus leads the brain to update its models, i.e. "an expected event does not need to be explicitly represented or communicated to higher cortical areas which have processed all of its relevant features prior to its occurrence" (Bubic et al. 2010, 10). On a PEM approach, when the predictions generated by low-level models are adequate, there is no need for the brain to engage in complex computation in order to make sense of incoming sensory information.

Another issue that is frequently brought up in connection with mindreading is the frame problem, which arises from the fact that there appears to be no intrinsic limit on the scope of the information that is relevant to mindreading processes (McCarthy and Hayes 1969; Heal 1997; Wilkerson 2001). For TT, the problem is how to decide which theoretical principle one should apply in a particular situation, given that the information that is potentially relevant to this decision is in principle unlimited and could come from any domain. For ST, the problem is how we are able to identify the mental states of the other agent, when the information that is potentially relevant to this identification is in principle unlimited and could come from any domain (Spaulding 2010; Gallagher 2012).

To see why this is not a problem for PEM, let us briefly consider Kilner et al. (2007) account of the mirror neuron system.[9] Because of the special visuo-motor properties of mirror neurons systems, proponents of ST have been attracted to the idea that mirror neurons enable us to understand other people's intentions simply by observing and covertly simulating their actions (Rizzolatti and Craighero 2004; Gallese et al. 2004). One way of expressing this idea is that mirror neurons are part of a larger system that transforms low-level representations of (observed) movement kinematics into high-level representations of action intentions. But how might this work? Take the example of the red teacup again. When I reach out in order to grasp it, my brain attempts to predict the sensory consequences of my action. It does so by means of a forward model, which runs a simulation of the movement kinematics on the basis of my action intention (Wolpert et al. 1995, 2003). But when I observe another agent reaching out for the cup, this sequence needs to be reversed. In order to achieve this, one possibility is that the brain generates an inverse model, which infers the intention of the agent from the observed movement kinematics. As Kilner et al. (2007) point out, however, the problem is that such a model assumes a one-to-one relation between the observed movement kinematics and its cause, when in general the relation is many-to-many. On the one hand, the same movement kinematics can be caused by different action intentions: an agent flipping a light switch may have the intention of turning the light either on or off, depending on the context. On the other hand, the same intention may issue in different movement kinematics: an agent might turn the light on by flipping the switch with her index finger or with her middle finger, or, if her hands are occupied, with her knee, etc.[10]

[9] The mirror neuron system is a network of brain regions (i.e. in ventral premotor cortex, inferior parietal lobe, and somatosensory areas) that is active either when an agent performs an action or when that agent observes another agent performing the same or a similar action (Frith and Singer 2008; Kilner et al. 2007; Michael 2011).

[10] See also Jacob and Jeannerod (2005)'s thought experiment, in which Dr. Jekyll may exhibit the same motor kinematics in curing his patient as Mr. Hyde does in torturing his victim. In order to differentiate between these two scenarios, it is necessary to draw upon some information beyond mere kinematics.

What we are facing here is basically ST's problem of determining relevance. However, Kilner et al. (2007) show that this problem doesn't arise if we understand the mirror neuron system as a hierarchical model operating in accordance with the principles of PEM. On this view, our brains do not need to start from scratch each time we observe someone performing an action—they arrive at the scene with prior expectations about how action intentions are causally related to movement kinematics. These prior expectations constrain the hypothesis space, and enable the brain to home in on likely interpretations of others' actions that can be tested and updated against the ongoing observation of behavior. Thus, when we observe how another agent flicks the light switch, we expect it will turn the lights on (given a certain context, e.g., during the evening or in a dark room), and on the basis of certain priors we expect some of the fingers to be involved—although we do not know precisely which ones, and we might be mistaken.

At this point it is interesting to reconsider Goldman's (2006) hybrid ST/TT account, which was (at least partly) motivated by the attempt to solve ST's frame problem. Goldman suggested that 'pure simulationism' is insufficient insofar as "[t]heorizing seems necessary to generate hypotheses about states responsible for the observed effects, hypotheses presumably prompted by background information." (p. 45) According to him, mindreading is executed at the personal level by simulation, which is in turn implemented at the sub-personal level by an underlying theory. "How could simulation be executed unless an algorithm for its execution is tacitly represented at some level in the brain? Isn't such an algorithm a sort of theory?" (ibid.).

Indeed, such an algorithm is precisely what PEM provides with Bayes' theorem. Although critics of TT have raised the question of what it means to employ theoretical principles at the sub-personal level (e.g., Gallagher 2001, 2008), PEM seems to offer a straightforward solution to this problem. For its core idea is that all perception and action is based upon unconscious inference. This suggests that social cognition is *theoretical* in the sense that, in order to predict other agents' future behavior, the brain draws inferences about the hidden mental causes of that behavior. But does it also mean that simulation is involved? The key claim of ST is that we predict and/or explain the behavior of another agent by using our own decision-making processes as a *physical simulation* of his or her decision processes. If this is how we should interpret ST, then it seems that PEM does not *necessarily* involve simulation: hierarchical models of other agents need not be based on the models that inform our own decision-making capacities (except in the trivial sense that all processes implement Bayesian algorithms). They might be based on our prior experience with the actions of a certain type of agent in a certain type of situation, instead of a simulation of how we ourselves would act in that situation. On the other hand, as Kilner et al. (2007) have shown, it remains an empirical possibility that the models of other agents that are generated by the brain are in fact based on models of ourselves. In that case, the models need to be sufficiently similar to allow for the application of one's own models to understand the other's behavior. As Friston and Frith (2015, p. 401) argue, "internal or generative models used to infer one's own behaviour can be deployed to infer the beliefs (e.g., intentions) of another—provided both parties have sufficiently similar generative models." (see Quadt 2017 for discussion).

Let us conclude this section with a general observation about the place of mindreading in the PEM framework. In anticipating other people's behavior, it is important not only to keep track of their mental states at a given time but also to draw upon information about features obtaining across longer time scales—i.e. personality traits and reputation, such as the degree to which they are trustworthy, lazy, prone to temper tantrums, etc. (Moutoussis et al. 2014). For example, there is evidence that people use information about other people's reputation (i.e. gossip) in order to generate expectations about how cooperative those people will be in economic games, and to calibrate their own behavior accordingly (Sommerfeld et al. 2007). What this shows is that PEM expands the scope of interest beyond mindreading to other phenomena that are also crucially important for social cognition. Indeed, mindreading does not appear to be so much a keystone of social cognition within this framework as one of many components that enhance the more basic domain-general ability to predict events in the environment that may or may not involve people.[11]

6 Situating PEM in Contemporary Debates on Social Cognition: Embodied Cognition

Recent embodied cognition approaches to social cognition (e.g., Gallagher 2001, 2012; Hutto 2004, 2008; Ratcliffe 2005, 2007; Zahavi 2004) have argued that our understanding of others is essentially: (1) embodied, i.e. shaped and structured by intra- and extra-cranial (bodily) processes, (2) enactive, i.e. not (primarily) about passively representing the mental states of others, but rather about dynamically interacting with them, and (3) embedded or 'situated' in a broader social and pragmatic context and based on a history of interactions, which constrains the information that is relevant to our understanding of them.

To what extent are these insights compatible with the main principles of PEM? Let us first consider the third claim that social cognition is embedded. PEM agrees that understanding others is situated and based on a history of interactions, insofar as it maintains that the predictions made by our brains are contextualized and based on prior expectations (Clark 2013, 2016). In fact, as we have argued in the previous section, this is how it is able to avoid the frame-problem. Now, given the importance accorded to prior expectations by the PEM framework, it is crucial to explain where those expectations come from. This is where embodied cognition approaches could actually be of help. Since our brains have evolved to make predictions that are specifically tailored to the bodies in which they reside and to the environments in which those bodies live, it only seems to make sense to retain a perspective which gives pride of place to bodies and environments As Clark puts it, to identify the structure of the space of priors that is specific to the human species, a "deep (but satisfyingly

[11] In fact, given that representations within this framework are distributed across multiple hierarchical levels, mindreading itself can be seen to fracture into multiple levels of inference about other agents' minds, as illustrated by the discussion in Sect. 3.

natural) engagement with evolutionary, embodied, and situated approaches" is needed. (p. 200)

PEM is also compatible with the assumption that social cognition emerges as the result of dynamic interactive processes. However, PEM suggests that the relevant contrast is not between dynamic interaction and passive representation, but instead between two ways of reducing prediction errors: active inference and perceptual inference. Indeed, it is the notion of active inference that invites us to incorporate social *interaction* in an overall framework for understanding social cognition. Quadt (2017) uses the term 'interactive inference' to describe the minimization of prediction error in a social environment, which serves to disambiguate competing models about other agents. She distinguishes between two kinds of interactive inference strategies that we employ in our social engagements. In replicative active inference the bodily state (e.g., posture, movements) of another person is mimicked in order to supplement exteroceptive information about them with interoceptive and proprioceptive information. Mimicry, synchronization and automatic imitation are ways in which replicative active inference helps us to make our predictions of other agents more precise by increasing the number of signal sources that yield relevant information. Complementary interactive inference refers to changing one's internal or external environment in response to another person. This is done to regulate the other's current state (e.g., mothers who lower their bodily temperature to cool down the infant's feverish body; Nyqvist et al. 2010), or to evoke further behavioral responses that serve as additional exteroceptive input (e.g., using gestures to express one's uncertainty).

Social interaction not only provides input to individual brains but enables us to jointly monitor each other's actions and speech, communicate error signals to each other (Pickering and Garrod 2014), and also to communicate levels of confidence (i.e. higher-order predictions) to each other (Bahrami et al. 2010). These benefits make it possible to pool together cognitive resources in order to detect errors all the better, and to ensure that the common ground necessary for successful joint decision making and joint action is sustained (Bahrami et al. 2012). In fact, Shea et al. (2014) have recently proposed that adult humans' ability to explicitly monitor and evaluate their own cognitive processes has been shaped evolutionarily and culturally by the primary function of social coordination.

Insofar as active and perceptual inference can be viewed as different ways to reduce prediction error, they cannot be characterized as 'passive'. However, according to most proponents of PEM, both types of inference do depend on a robust notion of representation (but see Bruineberg et al. 2016 for an exception). And this is hard to reconcile with more radical anti-representationalist versions of embodied cognition (e.g., Ratcliffe 2007; Gallagher 2008; De Jaegher et al. 2010; Hutto and Myin 2013). Furthermore, the assumption that prediction error minimization relies on a robust notion of representation sometimes leads to a much weaker view of the importance of the body. For example, Hohwy (2016) has argued that although the concept of active inference entails a central role for the body in reducing prediction error, this role is only significant in the sense that the body is *represented* in the model, as a parameter useful for minimizing prediction error. Others, however, argue that PEM works with a different notion of representation, one that is

'action-oriented' rather than 'action-neutral', and one that does not "seem to imply the presence of inner models or content-bearing states of the kind imagined in traditional cognitive science. Instead, what are picked out seem to be physical processes defined over states that do not bear contents at all." (Clark 2013, p. 18; see Downey 2017; Dolega 2017 for further discussion on the role of representations).

In other words, the extent to which PEM is compatible with claims about the importance of embodiment and the role of representations (at least) in social cognition partly depends on the various commitments of proponents of embodied cognition and proponents of PEM. Unfortunately, a detailed discussion of these commitments falls outside the scope of the present paper. Instead we will focus on some challenges that the PEM approach to social cognition still needs to address, and point out some general directions for future research

7 Further Questions and Future Research

One of the biggest challenges for the PEM framework—in general, i.e. not only as it pertains specifically to social cognition—in the near future will be to find direct neuroscientific evidence for the assumption that the brain is actually organized in terms of a hierarchical predictive coding model.[12]

At the moment, there is no conclusive evidence for the existence of two functionally distinct sets of neurons (i.e. 'error units' and 'representations units'), nor for any particular hypothesis about the interaction between any such sets—although recent work by Bastos (2013; see also Bastos et al. 2012) is suggestive in this respect. The evidence that does exist falls into three categories: (1) evidence that PEM models do a better job of predicting neural responses (e.g. in EEG or imaging studies) than competitor models—e.g. Pinotsis et al. (2014)'s recent report of layer-specific evidence for precision optimization, a key element of hierarchical predictive coding (see also Spratling 2012; Seth 2013); (2) behavioral demonstrations of Bayesian-like performance—e.g., binocular rivalry (Hohwy et al. 2008); and (3) simulations of predictive coding strategies that explain a variety of observed effects—e.g., Rao and Ballard (1999)'s model of predictive coding in the visual cortex.[13] Crucially, only the first category can directly support any firm conclusions about neural mechanisms. Moving forward, then, it will be important to assess to what extent category-1 evidence is forthcoming.

As far as social cognition research is concerned, indirect evidence for the PEM approach has predominantly been reported in the third category. For example, Baker et al. (2011) have presented a Bayesian Theory of Mind framework to model subjects' joint inferences about another agents' desires and beliefs about unobserved aspects of the environment. One obvious strategy is to take these formal PEM models as a starting point and try to map them onto what is currently known about the

[12] See Jazayeri (2008) for a review of findings that may be interpreted as providing such evidence; but see Bowers and Davis (2012, pp. 404–405) for criticism of Jazayeri's conclusions.

[13] See Clark (2013) for a fuller assessment of this issue.

neural mechanisms underlying social cognition—in this case the so-called 'Theory of Mind network', which is usually taken to consist of the anterior cingulate cortex, the temporal-parietal junction, the superior temporal sulcus and the temporal poles (see Frith and Frith 2003; Amodio and Frith 2006). However, one challenge for this strategy is that the PEM approach is very likely to put different kinds of constraints on the interpretation of the existing neuroscientific data. For instance, if we take for granted that only unexpected input generates an error signal (which is one of the central assumptions of PEM), then we should hypothesize that only unexpected input causes brain activity in higher cortical regions. This 'explaining away the input' hypothesis is clearly a solid basis for the generation of hypotheses that would be testable by cognitive neuroscience, but it may also require modifications of existing paradigms. This makes it difficult to apply PEM directly to data from previous research on social cognition.

In order to refine the PEM approach as a framework for social cognition research, we propose a combination of the following three strategies: (a) to reconsider formal PEM models in light of existing neuroscientific and psychological findings in social cognition research; (b) to re-interpret existing data from social cognition research in terms of the specific constraints that follow from adopting the PEM approach; and (c) to refine existing experimental paradigms in order to implement specific hypotheses arising from the PEM framework.

To conclude, it remains to be seen whether or not PEM will prove to be more successful as a framework for social cognition research than its predecessors were—just as it remains to be seen whether PEM will be successful as a general theory of neural processing. For the moment, however, we hope at least to have demonstrated that it is well worth finding out.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The role of medial frontal cortex in social cognition. *Nature Reviews Neuroscience, 7,* 268–277.

Apperly, I. A. (2008). Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study ''theory of mind''. *Cognition, 107*(1), 266–283.

Apperly, I. A. (2011). *Mindreaders: The cognitive basis of "theory of mind"*. New York: Psychology Press.

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.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback in the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 3–8.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329,* 108141085.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14,* 110–118.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the thirty-third annual conference of the cognitive science society* (pp. 2469–2474).

Bastos, A. M. (2013). Ph.D. thesis, University of California Davis.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76,* 695–711.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47,* 129–141.

Bermúdez, J. L. (2003). The domain of folk psychology. In A. O'Hear (Ed.), *Minds and persons* (pp. 25–48). Cambridge: Cambridge University Press.

Blokpoel, M., Kwisthout, J., & van Rooij, I. (2012). When can predictive brains be truly Bayesian? *Frontiers in Theoretical and Philosophical Psychology, 3*(460), 1–3.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138,* 389–414.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*. https://doi.org/10.1007/s11229-016-1239-1.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience, 4,* 25. https://doi.org/10.3389/fnhum.2010.00025.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*, 337–342.

Butterfill, S., & Apperly, I. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637.

Carruthers, P. (2016). Two systems for mindreading? *Review of Philosophy and Psychology*, *7*(1), 141–162.

Casler, K., & Kelemen, D. (2005). Young children's rapid learning about artifacts. *Developmental Science, 8,* 472–480.

Christensen, W., & Michael, J. (2013). Review of Apperly, Ian. *Mindreaders: The Cognitive Basis of Theory of Mind Phenomenology and the Cognitive Sciences, 12*(4), 907–914.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Costantini, M., Galati, G., Ferretti, A., Caulo, M., Tartaro, A., Romani, G. L., et al. (2005). Neural systems underlying observation of humanly impossible movements: An FMRI study. *Cerebral Cortex, 15,* 1761–1767.

Dahl, A., Schuck, R. K., & Campos, J. J. (2013). Do young toddlers act on their social preferences? *Developmental Psychology, 49*(10), 1964–1970.

De Bruin, L. C., & Newen, A. (2012). An association-based account of false belief understanding. *Cognition, 123*(2), 240–259.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences, 14*(10), 441–447.

Dolega, K. (2017). Moderate predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Downey, A. (2017). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*. https://doi.org/10.1007/s11229-017-1442-8.

Dunfield, K. A., & Kuhlmeier, V. A. (2010). Intention-mediated selective helping in infancy. *Psychological Science, 21*(4), 523–527.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences, 99,* 9602–9605.

Fiebich, A. (2014). Mindreading with ease? Fluency and belief reasoning in 4- to 5-year-olds. *Synthese*, *191*(5), 929–944.

Friston, K. (2002). Functional integration and inference in the brain. *Progress in Neurobiology, 68,* 113–143.

Friston, K., & Frith, C. (2015). A duet for one. *Consciousness and Cognition, 36,* 390–405.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104*(1–2), 137–160.

Frith, C., & Singer, T. (2008). The role of social cognition in decision making. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *363*, 3875–3886.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 358,* 459–473.
Gale, A., Spratt, G., Chapman, A. J., & Smallbone, A. (1975). EEG correlates of eye contact and interpersonal distance. *Biological Psychology, 3*(4), 237–245.
Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies, 8,* 83–108.
Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition, 17,* 535–543.
Gallagher, S. (2012). In defense of phenomenological approaches to social cognition: Interacting with the critics. *Review of Philosophy and Psychology, 3*(2), 187–212.
Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*(9), 396–403.
Goldman, A. (2006). *Simulating minds: The philosophy, psychology and neuroscience of mindreading*. New York: Oxford University Press.
Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4,* 223–233.
Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry, 51,* 59–67.
Heal, J. (1997). Simulation, theory, and content. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind* (pp. 75–89). Cambridge: Cambridge University Press.
Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*(4), 167–170.
Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*. https://doi.org/10.1111/desc.1214.
Hietanen, J. K., Leppänen, J. M., Peltola, M. J., Linna-aho, K., & Ruuhiala, H. J. (2008). Seeing direct and averted gaze activates the approach-avoidance motivational brain systems. *Neuropsychologia, 46*(9), 2423–2430.
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259–285.
Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701.
Hood, B., Willen, J., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, *9*(2), 131–134.
Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.
Hutto, D. D. (2004). The limits of spectatorial folk psychology. *Mind and Language, 19,* 548–573.
Hutto, D. (2008). *Folk psychological narratives: The sociocultural basis of. understanding reasons*. Cambridge, MA: MIT Press.
Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences, 9*(1), 21–25.
Jazayeri, M. (2008). Probabilistic sensory recoding. *Current Opinion in Neurobiology, 18,* 431–437.
Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*(3), 307–321.
Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166.
Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.
McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence, 4,* 463–502.
Michael, J. (2011). Interactionism and mindreading. *Review of Philosophy and Psychology*, *2*(3), 559–578.
Michael, J., Sandberg, K., Skewes, J., Wolf, T., Blicher, J., Overgaard, M., et al. (2014). Continuous theta burst demonstrates a causal role of premotor homunculus in action interpretation. *Psychological Science, 25,* 963–972. https://doi.org/10.1177/0956797613520608.
Michael, J., Sebanz, N., & Knoblich, G. (2016). The sense of commitment: A minimal approach. *Frontiers in Psychology*, *6*, 1968.
Michael, J., & Székely, M. (2017). Goal slippage: A mechanism for spontaneous instrumental helping in infancy? *Topoi*. https://doi.org/10.1007/s11245-017-9485-5.

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). A formal model of interpersonal inference. *Frontiers in Human Neuroscience, 25*(8), 160.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41,* 269–280.

Nichols, K. A., & Champness, B. G. (1971). Eye gaze and the GSR. *Journal of Experimental Social Psychology, 7,* 623.

Nichols, S., & Stich, S. (2003). *Mindreading. An integrated account of pretence, self-awareness, and understanding of other minds*. Oxford: Clarendon Press.

Nyqvist, K. H., Anderson, G. C., Bergman, N., Cattaneo, A., Charpak, N., Davanzo, R., et al. (2010). Towards universal Kangaroo Mother Care: Recommendations and report from the First European conference and Seventh International Workshop on Kangaroo Mother Care. *Acta Paediatrica, 99*(6), 820–826.

Pelphrey, K., Singerman, J., Allison, T., & McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: The influence of context. *Neuropsychologia*, *41*(2), 156–170.

Pelphrey, K., & Van der Wyk, B. (2011). Functional and neural mechanisms for eye gaze processing. In A. Calder, G. Rhodes, M. Johnson, & J. Haxby (Eds.), *OUP Handbook of face perception* (pp. 591–604). Oxford: Oxford University Press.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308,* 214–216.

Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience, 8,* 132.

Pinotsis, D., Robinson, P., beim Graben, P., & Friston, K. (2014). Neural masses and fields: Modelling the dynamics of brain activity. *Frontiers in Computational Neuroscience, 8,* 149.

Quadt, L. (2017). Action-oriented predictive processing and social cognition. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Qureshi, A., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective-selection, not Level-1 visual perspective-calculation: Evidence from a dualtask study of adults. *Cognition, 117,* 230–236.

Rakoczy, H. (2008). Taking fiction seriously: Young children understand the normative structure of joint pretend games. *Developmental Psychology, 44*(4), 1195–1201.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology, 44*(3), 875–881.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2*(1), 79–87.

Ratcliffe, M. (2005). Folk psychology and the biological basis of intersubjectivity. In A. O'Hear (Ed.), *Philosophy, biology and life. Royal Institute of Philosophy Supplement* (Vol. 56, pp. 211–233). Cambridge: Cambridge University Press.

Ratcliffe, M. (2007). *Rethinking commonsense psychology: A critique of folk psychology, theory of mind and simulation*. Basingstoke: Palgrave Macmillan.

Rizzolatti, G., & Craighero, L. (2004). The mirror neuron system. *Annual Review of Neuroscience, 27,* 169–192.

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

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2013). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 929–937.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55,* 87–124.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience, 7,* 413–422.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology*, *18*(9), 668–671.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 656–663.

Shea, N. J., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18,* 186–193.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Science USA, 104,* 17435.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition, 130,* 1–10.

Spaulding, S. (2010). Embodied cognition and mindreading. *Mind and Language, 25*(1), 119–140.

Spratling, M. W. (2012). Predictive coding accounts for V1 response properties recorded using reverse correlation. *Biological Cybernetics, 106*(1), 37–49.

Tamietto, M., Castelli, L., Vighetti, S., Perozzo, P., Geminiani, G., Weiskrantz, L., et al. (2009). Unseen facial and bodily expressions trigger fast emotional reactions. *Proceedings of the National Academy of Sciences of the United States of America, 106*(42), 17661–17666.

Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., & Davis, G. (2009). Social cognition modulates the sensory coding of observed gaze direction. *Current Biology, 19,* 1274–1277.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307,* 1642–1645.

Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development, 81*(6), 1661–1669.

Van der Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. A. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science, 20*(6), 771–777.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72,* 584–655.

Wilkerson, W. S. (2001). Simulation, theory, and the frame problem. *Philosophical Psychology*, *14*(2), 141–153.

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.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London, Series B, Biological sciences, 358*(1431), 593–602.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269,* 1880–1882.

Zahavi, D. (2004). The embodied self-awareness of the infant: A challenge to the theory-theory of mind? In D. Zahavi, T. Grünbaum, & J. Parnas (Eds.), *The structure and development of self-consciousness: Interdisciplinary perspectives* (pp. 35–63). Amsterdam: John Benjamins.