Trends in Cognitive Sciences

CellPress
REVIEWS

Opinion

Is There a 'Social' Brain? Implementations and Algorithms

Patricia L. Lockwood [1,2,3,*] Matthew A.J. Apps,[1,2,3,7] and Steve W.C. Chang[4,5,6,7]

A fundamental question in psychology and neuroscience is the extent to which cognitive and neural processes are specialised for social behaviour, or are shared with other 'non-social' cognitive, perceptual, and motor faculties. Here we apply the influential framework of Marr (1982) across research in humans, monkeys, and rodents to propose that information processing can be understood as 'social' or 'non-social' at different levels. We argue that processes can be socially specialised at the implementational and/or the algorithmic level, and that changing the goal of social behaviour can also change social specificity. This framework could provide important new insights into the nature of social behaviour across species, facilitate greater integration, and inspire novel theoretical and empirical approaches.

Highlights

A central question in psychology and neuroscience is the extent to which social behaviour is subserved by dedicated processes or systems that are 'socially specific' or shared with other 'non-social' cognitive, perceptual, and motor faculties.

We suggest that a process can be socially specific at different levels of explanation. This approach could help to clarify the role of mirror neurons in social contexts and whether social learning is uniquely 'social'. Experimental design should be guided by an appreciation that social specificity is possible at different levels.

Examining social behaviour across species can give unique clues about different implementations and algorithms. For example, converging evidence highlights the anterior cingulate gyrus as crucial for processing social implementations, and that 'theory of mind' is a putative social algorithm.

Social Specificity through the Lens of Marr

Many behaviours occur in a social context. Social behaviours, in some form, are exhibited across a surprisingly broad array of species from single-celled microorganisms [1] to rodents [2], fish [3], and primates [4]. However, a core question for psychology and neuroscience is whether there are cognitive processes, brain areas, circuits, or cells that process information in a manner that is socially specific. That is, are there processes that come online only in social situations in a way that is somehow different to what is required for 'non-social' cognitive, motor, and perceptual abilities?

We draw on the pioneering idea of **Marr's levels** (see Glossary) [5] to provide a framework to understand and test whether cognitive and neural processes are socially specialised or not. We argue that a process may be considered 'social' at the algorithmic level – it encodes a specific algorithm or rule that is different from what is being processed in a non-social domain – and/or at the implementational level, the same algorithm is used, but it is processed in a different brain area, circuit, or cell. Moreover, we suggest that changing the social goal of the information-processing system (computational level), such as during cooperation or competition, can change social specialisation at the other levels. These levels of description are often overlooked when studying information processing in social contexts. We contend that this can lead to inaccurate conclusions about whether cognitive or neural processes are specialised, and call for a more nuanced approach to the phrase 'the social brain' beyond its simple connotation.

Marr's Framework: Computation, Algorithm, and Implementation

Marr's framework [5] argued that, to understand an information-processing system, it is crucial to consider multiple levels of explanation – computational, algorithmic, and implementational (Figure 1) [6]. The highest level of description, computational, describes the 'why' of a system or the goal that it intends to perform. For example, if we want to understand bird flight we cannot do so 'by only studying the feathers' [5]. We first need to know that the goal of the bird is to fly. The second level is the algorithmic – 'what' rules does the brain apply for a particular operation? This would be the bird's flapping of its wings. The final level, implementational, is 'how' the brain achieves a particular operation. For a bird, this would be its feathers.

[1]Department of Experimental Psychology, University of Oxford, Oxford, UK
[2]Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK
[3]Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK
[4]Department of Psychology, Yale University, New Haven, CT, USA
[5]Department of Neuroscience, Yale University School of Medicine, New Haven, CT, USA
[6]Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, CT, USA
[7]Equal contributions

*Correspondence:
patricia.lockwood@psy.ox.ac.uk
(P.L. Lockwood).

 https://doi.org/10.1016/j.tics.2020.06.011

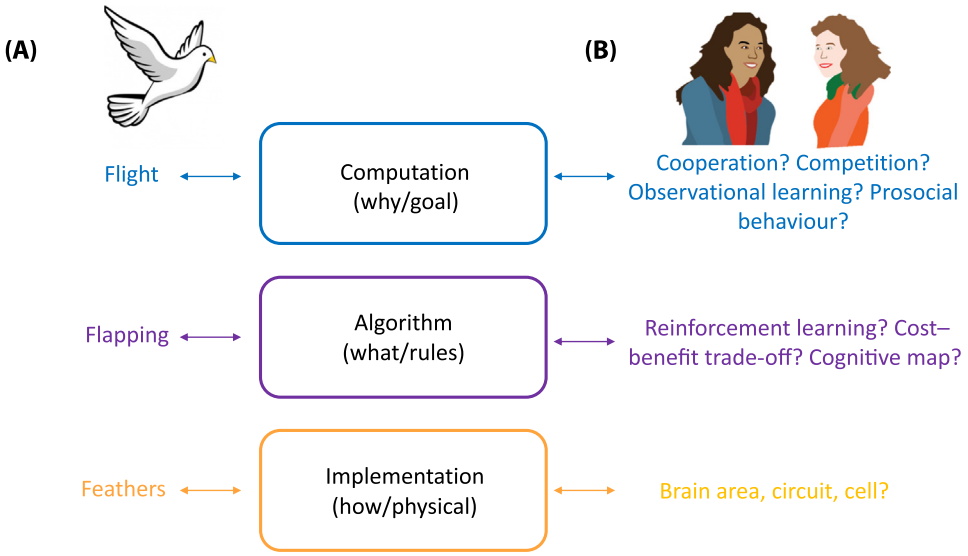

Figure 1. Marr's Three Levels of Analysis for Non-Social and Social Behaviour. (A) If we want to understand the goal of a bird to fly, we cannot simply study its feathers. We need to know that the bird's goal is to fly (computation), which it does by flapping its wings (algorithm), where the aerodynamics of flying depend on the feathers (implementation). (B) We argue that to understand how specific social behaviour is, as compared to other 'non-social' processes such as bird flight, we need to understand the social goal (are we cooperating, learning from, or helping the other person or group). Next, we need to understand the algorithm by which we achieve this. The relatively recent use of computational models such as reinforcement learning, cost–benefit trade-off, and cognitive maps are some examples of algorithms that could be used. Finally, we need to know how the social process is implemented, and in which brain areas, circuit, or cell it is realised. Crucially, we argue that a dissociation is needed between a social and non-social process, either at the level of algorithm or at the level of implementation, to conclude that there is social specificity. When designing experiments to test social specificity we should be looking for dissociations at algorithm or implementation.

How can this apply to social behaviour? The computational goal of social interactions is dictated by the nature and the intentions of the agent, such as cooperating, affiliating, or competing with conspecifics. The algorithmic level would be a particular, formalised, model of a social or cognitive process that is deployed only when engaging in a social interaction. Lastly, the implementational level would be the specific brain region, circuit, or cell in which the social process is realised. Although the number of levels and their independence are debated [6,7], Marr's theory provides an important organising framework suggesting that social specificity can be delineated at different levels. The most crucial point is that, for a process to be considered socially specific, there must be a dissociation between social and non-social processing at either the algorithmic or implemental levels, and alternative, similar, domain-general processes must be ruled out.

Marr's framework offers new insights into several debates in social neuroscience. We consider two notable examples. **Mirror neurons** or 'common currency' accounts suggest that social information is encoded based on an overlap in implementation (the same neuron fires similarly, or there is the same fMRI blood oxygen level-dependent, BOLD, response) to first- and third-person events. Examples include pain to self and another [8,9], monetary and **social reward** [10,11], and one's own or another's action goals [12,13]. This overlap is interpreted as a 'common coding' of different processes, namely that both first- and third-person understanding, or empathy, has occurred [6]. However, such conclusions about mirroring are drawn with reference only to the implementational level. What if a different algorithm is being used to

Glossary

Marr's levels: David Marr suggested that there were three levels of explanation for an information-processing system. The highest level is the computational or goal of the information-processing system. The second level is the algorithmic or the rules that the system applies. The third level is the implementational, or the physical realisation of the system.
Metacognition: the ability to attribute mental states such as beliefs, desires, and intentions to oneself.
Mirror neurons: neurons initially discovered in the monkey premotor cortex that fire similarly both when executing an action and when observing an action.
Optogenetics: a biological technique where light is used to control neurons that have been genetically modified to express light-sensitive ion channels. This technique is commonly used in rodent studies but it is not currently possible to safely use in humans.
Reinforcement learning (RL): learning associations between stimuli or actions and positive and negative outcomes. Learning is driven by how unexpected the outcome is.
Social pain: aversive nociceptive events that occur to other people.
Social prediction errors: differences between expected and actual outcomes involving others that occur during social interactions.
Social reward: rewards that are derived from, or obtained in the context of, social interaction.
Theory of mind: the ability to attribute mental states such as beliefs, desires, and intentions to other individuals.

understand the states of another? Without manipulating or controlling the algorithmic level, mirroring could reflect either a total absence of social specificity, an absence of social specificity only at implementation, or only in one particular circuit or cell. Indeed, it is clear that the goal level of the system is different when monitoring another's actions. We are not necessarily aiming to reproduce the actions ourselves, and therefore even if there is overlap at the implementation, there is likely to be functional dissociation. However, the other levels need to be measured using carefully controlled designs to manipulate them to understand what – if anything – the overlap means for social cognition and behaviour (Box 1).

Another major debate is whether social learning requires uniquely 'social' processes or arises from domain-general associative learning [14–16]. There is growing evidence that the same associative algorithms can indeed be used for both personal and social learning [4,17–20], which could be argued to reflect an absence of social specificity. However, what about the implementational level? In contrast to the algorithmic level, there is an increasing consensus for dissociation at the implementational level for some social learning processes in cells or circuits that are not involved in learning from the outcomes of one's own actions [17,19–25]. Thus, social learning may be 'socially specific' at the level of neural implementation, even though the algorithm may be the same.

These are only exemplars, but highlight how debates in social neuroscience and psychology can be addressed by considering which of Marr's levels one's debate addresses. Such an idea could inform experimental design in studies of social cognition and its neural basis. Key to addressing such debates empirically will be to use experimental designs in which one of the levels is held constant, either the algorithm or implementation, to test for specificity at the other level. For example, lesion and brain stimulation approaches can clearly examine the impact of disrupting

Box 1. The Importance of Non-Social Control Conditions

When designing an experiment to test whether a cognitive or neural process has a selective role in social behaviour, it is important to think about appropriate control conditions. This idea stems from basic principles in philosophy of science on the necessity of falsification [91]. For example, if we find that a particular brain area responds to both smiling faces and to monetary reward, can we conclude that this area is encoding something about how rewarding the stimulus is? We have not shown that this brain area is not involved in any other process, only that it is involved in two processes that share a common feature. Many studies in rodents and monkeys have included explicit 'non-social' control conditions [21,42,45,47,50,51]. In studies of self and other reward processing in monkeys, a 'neither' reward condition has been added, where a reward is seen as being delivered to neither the monkey nor their conspecific [45,63]. In rodent studies of observational fear conditioning, a more typical classical conditioning condition, without any social context, has been employed to try to rule out a domain-general response to aversive events. In some studies of theory of mind processing, a 'computer' condition or a physical object condition has been introduced to try to show specificity for theory of mind processing [65,78–80].

However, these control conditions are not always part of the experimental design. Sometimes it can be very difficult to create an equally matched non-social control that shares all or most attributes of a social stimulus except for its sociality – that is, a stimulus that is about or for another person or group. In the example of the computer condition, it may be that people anthropomorphise the computer and therefore still associate the stimulus with a social context, and it can be worth checking how participants perceive the condition. It is also well known that simple geometric shapes moving in a way that implies social interaction can be interpreted as social [92]. Therefore, a central factor in creating a social versus non-social condition appears to be the beliefs that the person has about whether the stimulus is social or non-social, rather than necessarily the observed behaviour. This is clearly shown in a study by Stanley and colleagues [93] who used a 2 × 2 design to probe the role of beliefs and behaviour in perceiving stimuli as social. Participants observed dot-motion animations and were instructed that they were either from prerecorded human movement or that they were computer-generated. They also manipulated whether the dot display followed biologically plausible or implausible velocity profiles. Participants experienced interference from the display when they were told that it reflected human movement, regardless of whether the velocity profiles were biologically plausible or not. This study therefore supports the idea that inducing beliefs that a stimulus is social versus non-social is crucial for creating social and non-social experimental conditions.

the implementational level on the algorithmic level. We can disrupt a specific implementation and test whether a social or non-social algorithm is changed. At the algorithmic level, we might examine a common algorithm, such as a **reinforcement learning** (RL) process, and then test whether it is differentially implemented in a social (learning about rewards for others) versus non-social (learning about rewards for self) condition. Considering how to dissociate different levels of analysis is therefore important in generating experimental designs that aim to address specificity. Moreover, we suggest that it is critical that additional 'non-social' control conditions are tested, if identifying social specificity is the aim (Box 1). In the following sections we examine this hypothesis by using key examples from the field and across research in humans, non-human primates, and rodents.

Social and Non-Social Processing across Levels and Species

A first question to ask is – why might we expect to find social specificity at any level of explanation? Evolutionarily, animal species are adapted to physical environments, and, for species that often interact with their conspecifics, they are adapted to social environments too [26]. The social brain hypothesis argues that the cognitive abilities required for navigating through social environments shaped the large brains of primates relative to other animals [27–30], and the preceding social intelligence hypothesis argues that social group structures were an evolutionary pressure that drove the emergence of higher intelligence in animals [31,32]. In rodents, olfaction and vocalisations in social contexts are strongly linked to evolutionary fitness [33,34]. However, even if we consider that it is plausible that the brains of different species might be adapted to their social environments, it does not necessarily follow that neural systems and processes must be specialised.

Moreover, although it goes without saying that humans are social creatures, the complexity and boundaries of social behaviour in non-human primates, and particularly rodents, is widely debated [2,4,35,36]. For example, whereas many would agree that humans engage in social processing such as empathising with others and **theory of mind**, such processes are much more controversial in non-human primates and rodents [4,35–37]. This is important when examining social specificity – if the same social cognitions and behaviours are not shared across species, then we might expect the algorithms and implementations also to be different rather than conserved. In addition to clear differences in the complexity of social behaviour, there are also differences in homology of brain areas [38] and methodological approaches [2,36] (Box 2). These methodological approaches can also vary greatly in their experimental resolution, from **optogenetics** in single cells in rodents to whole-brain neuroimaging in humans, and therefore specificity definition can change as a function of resolution. However, despite these differences, parallels can be drawn, perhaps most readily in the domain of appetitive and aversive processing

Box 2. Opportunities and Challenges of Cross-Species Comparisons

There are many challenges when trying to conduct and interpret findings that come from different species. Differences in homologies of brain areas and circuits, differences in experimental design, and differences in the resolution of methods are some of the most crucial. Despite these challenges, comparisons can be drawn and are key if we want to ultimately understand human behaviour and its pathologies. When conducting studies with humans to investigate social behaviour, we have the advantage of inquiring about their thoughts and feelings and to study the system we are trying to understand. However, we ethically cannot cause focal lesions, and we cannot currently use very precise optogenetic methods for manipulating single cells. Conversely, in rodents we can use very precise optogenetic methods, but then might question how similar rodent behaviour is to human social behaviour, and we certainly cannot inquire about their thoughts and feelings. This highlights the importance of using multiple methods and across different species to study social behaviour, and for being aware of the opportunities and challenges of most meaningfully comparing them to derive converging knowledge. It can also be useful for inspiring new empirical approaches and areas of research focus in different species. For example, work in monkeys on the ACCg in social behaviour inspired several subsequent studies in humans that also confirmed the importance of ACCg in human and rodent social behaviour [21,42]. The organising framework of Marr that suggests social specificity at the level of algorithm and implementation can be readily applied in different species and may allow us to draw greater connections between findings at multiple levels in future research.

because similar experimental techniques have been used across species [4,39,40]. As such, the following questions can be raised – are there specific neural circuits and cells that support social behaviours? Are they apparent at different levels of explanation? and is there evidence for clear dissociations in social specificity at the implementational and algorithmic levels?

Socially Specific Implementation?

Starting with the first question, of whether there are specific neural systems for social behaviour at the implementational level, the anterior cingulate cortex (ACC) is a key candidate (Figure 2). The ACC is also ideal for cross-species comparisons because it is relatively well studied in monkeys and rodents [21,38,41–44]. Most intriguingly, evidence points to important divisions in social and non-social implementation between subregions within the ACC, particularly the sulcus (ACCs) and gyrus (ACCg) [4,21,39,42,45,46] (Figure 2). A seminal study by Chang *et al.* [45] showed that there are varying levels of specialisation for social and self-oriented reward processing in ACCg and ACCs, respectively. The researchers recorded single-unit activity from ACCg, ACCs, and orbitofrontal cortex (OFC) while monkeys made decisions to deliver rewards to themselves, another monkey, or neither. OFC neurons predominately responded to received reward outcomes of self, and ACCs neurons tracked foregone reward outcomes of self. By contrast, neurons in the ACCg predominantly encoded the received reward outcome of a conspecific monkey, and some neurons responded to others' rewards exclusively (other-referenced) whereas another cluster showed a response that was 'mirror-like', encoding rewards of self and other (both-referenced). These results suggest that, at the level of single cells, there is some social specificity because more cells responded to others' rewards in ACCg than in the other areas tested.

Although some cells responded to the monkeys' own reward in ACCg, lesion studies can provide causal evidence for whether a region's function underlies social behaviour. Strikingly, when the

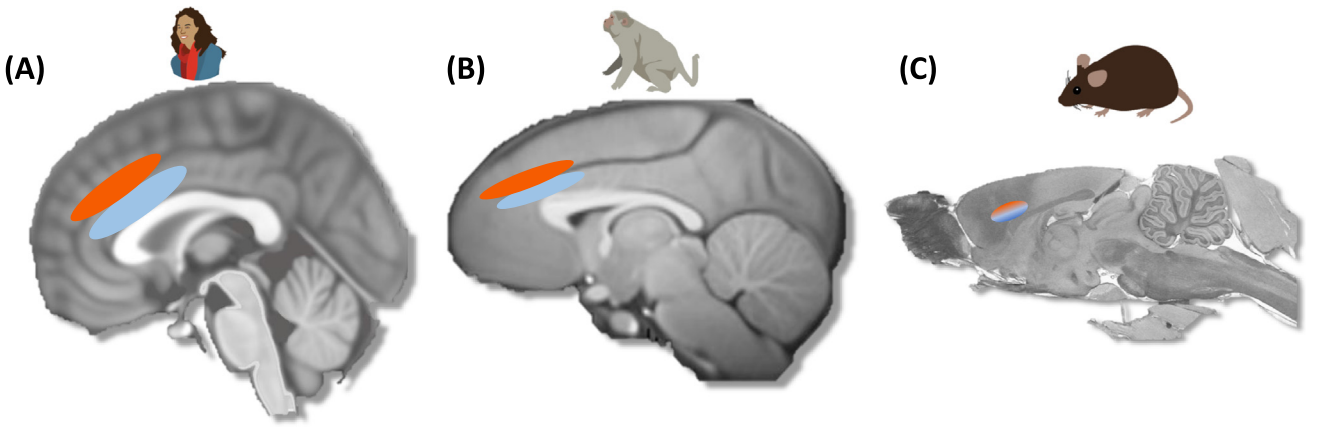

Figure 2. The Anterior Cingulate Cortex (ACC) and Socially Specific Implementation. (A) In humans, several studies have found that the gyral portion of the ACC (ACCg, light blue) responds to prediction errors and rewards predicted or delivered to other people, whereas the neighbouring sulcus (ACCs, red) responds to self-relevant reward signals and prediction errors [4,17,21,42,43]. Importantly, there is also evidence that ACCs responds to prediction errors and forgone rewards, hinting that this area might process a domain-general response to rewards not delivered to oneself instead of processing a socially specific signal [4,17,21,42,43]. (B) Converging evidence from focal lesion and single-unit recording studies suggests that the ACCg responds to other monkeys' rewards in terms of attention allocation and behavioural choice, where monkeys show a preference to reward others over rewarding neither self nor another. Instead, a large proportion of neurons in ACCs signal both self reward and 'neither' reward, consistent with an involvement in foregone rewards [45,47,48]. (C) In rodents, clear divisions of the sulcus and gyrus are not as readily apparent as they are in humans and monkeys, but roughly correspond to the areas known as cg1 and cg2. Evidence suggests that rodent ACC contains neurons that respond both to foot shocks delivered to the rat themselves and to the observation of shocks given to another rat. Importantly, these same neurons do not respond to fear conditioning [50,51]. It is an open question whether the response profile in rat ACC reflects the differences in homology between rodents and other species, or whether it reflects differences in brain evolution.

whole ACCg region is disrupted, attention to social information is impaired, whereas lesions to the neighbouring sulcus leave attention to social information intact [47]. Further corroborating the social specificity of these effects, ACCg lesions did not change the processing of emotional stimuli (a snake) or of control objects. Likewise, a recent study that lesioned the whole ACC found specific disruption of learning which stimuli rewarded others, but not self [48]. By contrast, preferences to reward self and other over neither with previously learned stimuli were preserved. It remains to be tested whether these effects were largely driven by the ACCg lesion [49]. However, they provide clear evidence that damage at the implementational level, to the ACC, selectively affects social learning (Figure 2).

Research in rodents has also pointed to a key role for the ACC in social behaviour [44,50–54]. It is thought that Cg1 and Cg2 may be a homologue of human ACC, although a division between the sulcal and gyral portions is not apparent in rodents as clearly as in humans and primates [38]. Converging evidence points to rodent ACC being linked to processing rewards and pain in social contexts with some specificity. For example, studies in rats have suggested that they avoid actions that harm others, and this effect is abolished by ACC inactivation [51]. Similarly, ACC inactivation disrupts observational fear learning while leaving classical conditioning intact [55]. By contrast, amygdala lesions disrupt both observational learning and classical fear learning, suggesting a role in learning *per se*. Dovetailing with the work in macaques [45], a larger proportion of neurons in rodent ACC respond specifically to other's pain (27%) and to both other's and self pain (34%) than to self pain alone (12%) [50]. Moreover, specific deactivation of the ACC region disrupts freezing responses in the **social pain** context but not to a non-social fear-conditioned sound. The inclusion and comparison of the fear condition in this study [50], and the 'neither' reward condition in the study with macaques [45], provides an important control for the 'both' condition (Figure 3) when establishing the nature of overlap to self and other at the implementational level. Establishing whether a neuron or brain area is specifically involved in 'pain' or 'reward' requires excluding that it is involved in general aversive and appetitive processing (Box 1).

By using paradigms with non-social controls, and by using both neural recordings and causal methods, these studies provide one of the clearest cases for social specificity at the level of implementation (Figure 3); there are cells in the ACC that specifically respond to other's pain and reward across humans, primates, and rodents, and ACC damage selectively disrupts social information processing.

Different Implementation but Same Algorithm?

The development of model-based fMRI in humans has seen many studies testing whether different competing algorithms can explain behaviour and map onto functional anatomy [56–58]. RL is perhaps the best exemplar of a clear algorithmic process, and has been applied extensively to understand self-relevant and social behaviour. RL describes how actions are associated with outcomes based on the unexpectedness (the 'prediction errors') and the valence of outcomes, quantifying how behaviours are positively or negatively reinforced by rewards or punishment [18,57,58]. In humans, evidence suggests that the same region – the ACCg – which putatively shows relatively high levels of social specificity for implementation, may do so under a common RL algorithm [17,21,23,24,39,59,60]. This includes comparing self-relevant learning to learning to associate stimuli with others [23], learning whether to trust advice from others [17], and teaching others what to choose [24] where specific **social prediction errors** drive learning. This pattern can be contrasted with other brain areas such as ventral striatum that have been repeatedly related to tracking RL signals but without any socially specific implementation. For example, several studies have shown that social learning during observation, prosocial behaviour, trait

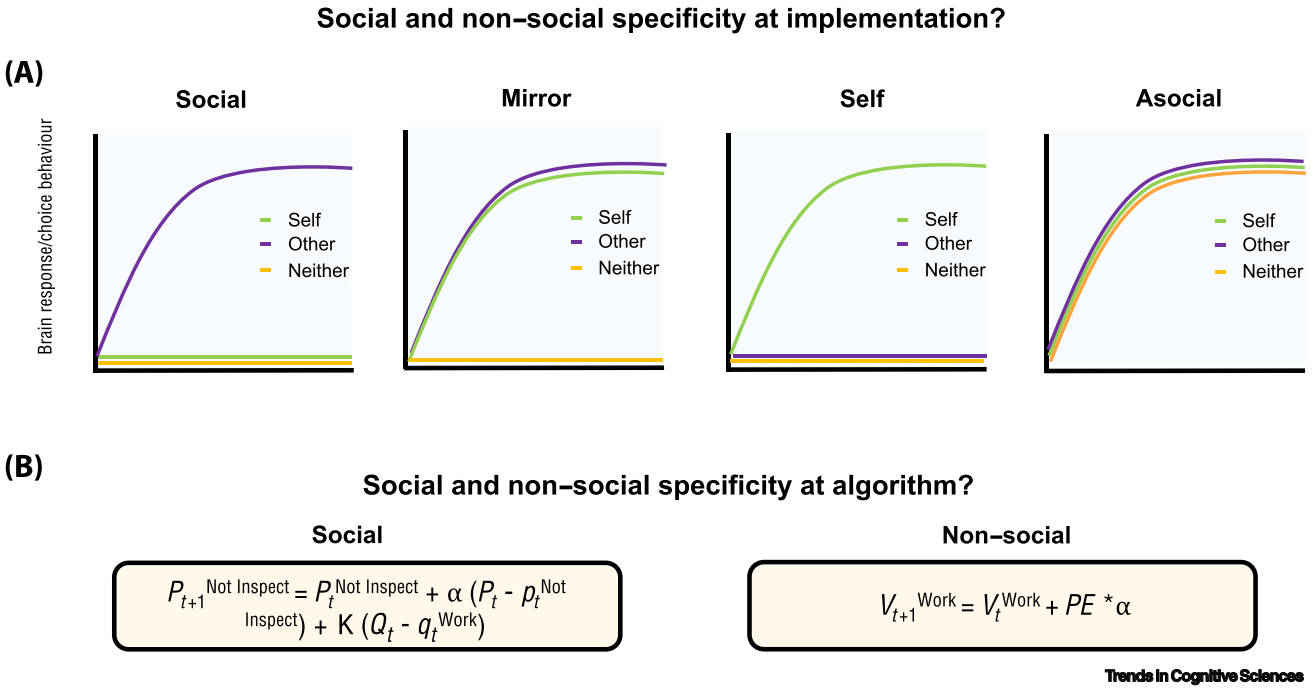

Figure 3. Social Specificity at the Implementational and Algorithmic Levels. (A) Schematic illustrating the hypothesised pattern of choice behaviour and/or brain response for socially specific information processing, mirror processing, self processing, or asocial processing at implementation. These patterns highlight how including the 'self' condition and/or an 'asocial' condition can help to show how socially specific a particular process is, and the types of brain or behavioural profile we would expect to see in each condition. (B) Examples of a 'social' and 'non-social' algorithm from Hill *et al.* [71] and Hampton *et al.* [72]. These algorithms can be distinguished in a task where participants are strategically interacting with another player either in the role of an 'employee' or an 'employer'. In the role of the employee, to maximize their payoff, participants must work when inspected and not work when not inspected. The authors compared different models to explain how participants played the game, a simple reinforcement learning model that simply tracked the reward outcomes regardless of the opponent (non-social) did not explain behaviour as well as an influence algorithm (social) that took into account the influence that the players' strategy has on the opponents behaviour. In the social example, the algorithm computes the decision of an agent at trial t as a function of both the agent's history of choice and the opponent's history of choice. This is contrasted to a non-social algorithm that simply computes the history of outcomes in the environment regardless of the opponent's history of choice. P_t is the opponent's probability of choosing an action (inspect or not inspect). α is the learning rate parameter. K is a constant that weights second-order beliefs and approximates the parameters of the opponent (the learning rate, temperature, and payoff matrix). Q_t is the employee's action at trial t, and qt is the employer's inferred probability that the employee will work. In the non-social example, Vt represents the action values that are updated based on the prediction error (PE) as to whether the action was selected and followed by reward or not. The prediction error is weighted by a learning rate (α). For further details see Hill *et al.* [71] and Hampton *et al.* [72].

understanding, trust learning, and non-social learning about rewards is commonly encoded in the ventral striatum [17,19,22,25,61,62]. Ventral striatum responses even seem to track prediction errors when no individual is associated with an outcome [22], consistent with the idea that, in humans, a domain-general learning algorithm may be implemented in the ventral striatum. By contrast, the evidence from the ACCg points to the possibility that brain areas can be socially specialised but may not need to implement a specialised algorithm.

Socially Specific Implementations beyond Cells and Areas

Methodological developments such as optogenetics, psychophysiological interactions, diffusion tensor imaging, and measures of synchrony allow for testing whether implementations can be socially specialised not only in single cells or brain areas but also in circuits. Several studies in non-human primates and rodents have hinted at socially specific circuit implementations. For example, the social specialisation of ACCg for social reward encoding extends to inter-regional coupling patterns in a ACCg–amygdala network [63], and projections between ACCg and basolateral amygdala (BLA) are specific for observational as compared to classical fear learning [52]. Activation of a BLA→mPFC pathway increases anxiety-like behaviours and

reduces social interaction, whereas inhibiting the same pathway increases social interaction and reduces anxiety [64]. This shows that, although research suggests a lack of social specificity when considering the whole amygdala [11], there could still be socially specific circuit-level implementations when interacting with medial prefrontal areas. In humans, fewer studies have directly compared social versus non-social connectivity, but notable examples include the ACCg–rostral cingulate connectivity that is present only when another's unexpected outcome is processed but not one's own [65], and a common associative neural network preferentially connects with the temporoparietal junction (TPJ) during social compared to direct fear learning [20]. These are only some examples of social specificity implemented in brain circuits.

Socially Specific Algorithms?

So far, we have discussed potential social specificity at the implementational level of cells, regions, and circuits. However, are there algorithms that are socially specific? This question is more challenging in terms of clearly defining algorithms, and there is wide controversy regarding the use of complex algorithms across species [4,66]. Nevertheless, it has long been argued that some social processes, such as theory of mind or 'mentalising' [67–70], may be socially specific – and thus, we contend, rely on specialised algorithms (Figure 3).

Several lines of research have begun to develop algorithms to model theory of mind processing, and explain behaviour in two-person social exchanges [4,10,40]. Although many of these models have been derived from those employed to explain economic preferences or more standard RL, they are clearly distinct and more sophisticated [71–73], and on the surface it is unclear why an agent would need such sophistication outside social interactions. For example, several studies [70–72] have required participants to play the role of an 'employer' or 'employee' where they choose to 'work' or 'shirk' (in the role of an employee), and interacted with the 'employer' who could inspect or not inspect what they were doing (Figure 3B). For the employee, rewards were maximised if they 'shirked' when not inspected and worked when inspected. The algorithm that best explained participants' behaviour took into account the influence that the employee's own actions would have on the employer's behaviour, and this was a better predictor than a simple RL that only took into account the history of outcomes.

Other studies have directly compared a socially specific framing (hide and seek) to a non-social framing (gambling) and showed that participants win more against mentalising agents, pointing to an 'added-value' of using mentalising when learning in social interactions [73]. The authors of this latter study also showed that non-human primate species do exhibit a precursor form of theory of mind algorithms [66], where they behave as if they were adjusting their estimate of others' likely responses to their own actions. In rodents, a putative precursor of theory of mind is much less apparent, and to our knowledge has not been tested. Rodents can exhibit RL, which would correspond to a very basic algorithm that could be used in social interaction and the most rudimentary theory of mind precursor, according to Devaine and colleagues [66]. It would be intriguing to test whether rodents can extend this basic associative process to learn associations that distinguish self from other, or to hold a concept of another animal that is not oneself, which would be necessary conditions for having theory of mind. It is plausible that rodents have this capacity given the aforementioned observational learning research showing socially specific modulation of electric shocks delivered to rodents themselves or to partner rodents [50,51]. There is also evidence in humans that simply forming associations between self, close others, and distant others is tracked in the TPJ, whereas ACCg specifically tracks learning links between stimuli and distant others [23]. Future studies could therefore probe further whether rodents demonstrably have a sophisticated concept of another animal, and the parts of the brain that are involved in that process.

Are the algorithms of theory of mind implemented in a socially specific manner? Several studies have suggested that the TPJ and dorsomedial prefrontal cortex (dmPFC) may process mentalising algorithms with social specificity [4,10,40,67,74]. In the two aforementioned work/shirk studies, the 'influence' algorithm was uniquely implemented in the dmPFC and TPJ [71,72]. However, a limitation of many theory of mind studies – in terms of assessing social specificity at the implementation level – is that they often lack a comparable 'self-relevant' or 'non-social' condition, making it difficult to conclude a specialised implementation (Box 1) [71,72,75]. As a result, there has been considerable debate as to whether processing in the dmPFC or TPJ is socially specialised, asocial [23], or reflects a common processing mechanism that is also engaged during self-monitoring and **metacognition** [76,77]. Other studies using non-social control conditions, such as 'computer' agents, have shown stronger responses to 'other' as compared to a computer [65,78–80] (see Box 1 for further discussion). Thus, two developments will be necessary to test the social specificity of theory of mind: (i) the use of appropriate non-social control conditions to test specificity at the implementational level, and (ii) the development of formal algorithms of other competitor processes, such as metacognition, that can be used to test for specificity at the algorithmic level.

A promising new direction for integrating across Marr's levels is the use of multivariate techniques, ranging from classifiers to more model-based representational similarity analysis (RSA) approaches. Such approaches have already been useful for showing that, within some regions, patterns of activity can be differentiated between self and other [81–85]. This includes physical pain and social rejection, as well as self and other valuations, in ACC [81], and patterns reflecting others' pain and other negative-valence stimuli, including disgust and unfair monetary exchange, in right anterior insula [83]. Simple classifiers decoding different self and other patterns may reflect distinct implementation – namely social specificity within a brain region. However, RSA techniques may bridge the gap between the algorithmic and implementational levels, and also test for differences at the algorithmic level. This is because the RSA approach inherently includes testing models for how information or stimuli are being represented – that is, how information is algorithmically structured [86]. Competing models can be generated that predict different algorithmic structures, and then the brain areas that correlate most strongly with them can be quantified. The approach can link the algorithmic and implementational levels by using the models to understand the brain imaging data, and, conversely, by using the brain imaging data to build the competing models [86].

Such an approach is also possible using parametric model-based imaging approaches that hypothesise a particular cognitive model, such as RL, to understand more about the implementational level in terms of function. Models therefore also provide a clearer link between the algorithmic and implementational levels than standard categorical analyses of brain data contrasting conditions, such as faces versus houses. Future research may be able to link levels of social specificity by using these multivariate and model-based techniques. These approaches can also shed light on the nature of the levels themselves, as well as on their interdependence, by highlighting the precise way in which they interact.

Does Changing the Goal Change Algorithms and Implementations?

The highest level of Marr's framework, the computational level, addresses the importance of the goals of an information-processing system. Across species, it is clear that goals of social behaviours – the social motivations – can differ from one context to the next. One minute we compete, the next we cooperate. However, can changing a goal modify social specificity at the other two levels? Although less work has directly tested this notion, there are hints that changing the social goal can indeed alter neural implementations. In rodents, a large proportion of ACC

neurons code the net value of rewards – the size of the reward discounted by the costs of competing with another for that resource. However, ACC neurons only code that net value when rodents were required to compete [87]. In monkeys, task outcome signals (i.e., winning or losing) in many lateral PFC neurons are gated by whether monkeys are competing for rewards or not [88]. In humans, whether we are cooperating or competing with others adjusts the extent to which the dmPFC tracks the performance of ourselves compared to others [89].

These findings, that support the ability of social goals to regulate specificity at the other levels, have potential implications for understanding disorders of social behaviour and their social uniqueness. For example, in group studies examining differences in neural implementation between patients and controls, it could be that differences in neural implementation or in the algorithms that are used between the two groups might appear to be algorithmic or implementational differences, when in fact it is the goal that is different between the groups and causes the changes in neural response. Evidence supporting this comes from one study that compared the neural responses of individuals with psychopathy and non-psychopathic participants while they viewed video clips of emotional hand interactions [90]. The authors found the group differences in neural responses were markedly reduced when the psychopathic offenders were instructed to empathise versus receiving no instructions. This study highlights how changing the social goal might change implementation, and the importance of matching motivation between groups when studying social specificity.

Concluding Remarks

Debates about social specificity have been at the core of social neuroscience and psychology for decades. We outline here how considering these questions within Marr's framework provides a novel perspective that may help to restructure discussions (see Outstanding Questions). Considering which of Marr's levels an experiment is testing at, and designing experiments that control and dissociate at one of the three levels, will allow us to reformulate questions across species. Utilising techniques that may help to bridge the gap between the algorithmic and computational levels, such as computational models of RL and economic decision-making, and multivariate techniques such as representational similarity analysis, will be important for moving forward. It is an open question how social specificity arises and what is conserved across species. Ultimately, the approach outlined here could help us to redefine the social brain by its implementations, algorithms, and computations.

Acknowledgments

This work was supported by a Medical Research Council Fellowship (MR/P014097/1), a Christ Church Junior Research Fellowship, and a Christ Church Research Centre grant to P.L.L.; a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/R010668/1) and a Wellcome Trust ISSF grant to M.A.J.A; and by grants from the National Institute of Mental Health (R01MH110750; R01MH120081) to S.W.C.C. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). We thank Colin Stanton for help with our illustrations.

References

1. Crespi, B.J. (2001) The evolution of social behavior in microorganisms. *Trends Ecol. Evol.* 16, 178–183
2. Chen, P. and Hong, W. (2018) Neural circuit mechanisms of social behavior. *Neuron* 98, 16–30
3. Bshary, R. *et al.* (2014) Social cognition in fishes. *Trends Cogn. Sci.* 18, 465–471
4. Wittmann, M.K. *et al.* (2018) Neural mechanisms of social cognition in primates. *Annu. Rev. Neurosci.* 41, 99–118
5. Marr, D. (1982) *Vision*, MIT Press
6. Krakauer, J.W. *et al.* (2017) Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490
7. Bickle, J. (2015) Marr and reductionism. *Top. Cogn. Sci.* 7, 299–311
8. Lamm, C. *et al.* (2011) Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage* 54, 2492–2502
9. Fan, Y. *et al.* (2011) Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neurosci. Biobehav. Rev.* 35, 903–911
10. Ruff, C.C. and Fehr, E. (2014) The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* 15, 549–562
11. Munuera, J. *et al.* (2018) Shared neural coding for social hierarchy and reward value in primate amygdala. *Nat. Neurosci.* 21, 415–423
12. Umilta, M.A. *et al.* (2001) I know what you are doing: a neurophysiological study. *Neuron* 31, 155–165
13. Chong, T.T.-J. *et al.* (2008) fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Curr. Biol.* 18, 1576–1580
14. Heyes, C. (2012) What's social about social learning? *J. Comp. Psychol.* 126, 193–202

Outstanding Questions

How do the algorithmic and implementational levels interact, and can multivariate and computational modelling approaches be used to bridge the gap?

Are there species differences in algorithm or implementation? It is possible that more specialised social processing might be carried out in evolutionarily ancient subcortical structures in non-primate animals compared to primates such as humans and monkeys. The evidence from non-human animals mostly reports elegant social specificity at the implementational level. It would be informative to also test for different algorithms, and an open question remains as to whether non-humans can use socially specific algorithms. This endeavour can be facilitated by recent advances in using computational models with clear algorithms to understand social behaviour.

Are socially specific algorithms and implementations innate and predetermined, or can they arise via associative learning?

How should we design future studies aimed at establishing the social specificity of a cognitive or neural process? We suggest that it is important to hold one level (computational, algorithm, or implementation) constant while examining the impact on another level.

Are there any brain areas, circuits, or cells that are uniquely socially specific? The strongest evidence seems to be for the ACCg. Are there algorithms that are socially specific? The strongest evidence suggests theory of mind processing.

Is the social versus non-social distinction in the brain categorical, or could there be a continuous relationship? The answers to this question can help to better understand how algorithmic, computational, or implementational levels were repurposed for social functions in brain evolution.

15. Cook, R. *et al.* (2014) Mirror neurons: from origin to function. *Behav. Brain Sci.* 37, 177–192
16. Catmur, C. *et al.* (2009) Associative sequence learning: the role of experience in the development of imitation and the mirror system. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 2369–2380
17. Behrens, T.E.J. *et al.* (2008) Associative learning of social value. *Nature* 456, 245–249
18. Lockwood, P.L. and Klein-Flügge, M. (2020) Computational modelling of social cognition and behaviour – a reinforcement learning primer. *Soc. Cogn. Affect. Neurosci.* Published online March 30, 2020. https://doi.org/10.1093/scan/nsaa040
19. Olsson, A. *et al.* (2020) The neural and computational systems of social learning. *Nat. Rev. Neurosci.* 21, 197–212
20. Lindström, B. *et al.* (2018) A common neural network differentially mediates direct and social fear learning. *NeuroImage* 167, 121–129
21. Apps, M.A.J. *et al.* (2016) The anterior cingulate gyrus and social cognition: tracking the motivation of others. *Neuron* 90, 692–707
22. Lockwood, P.L. *et al.* (2016) Neurocomputational mechanisms of prosocial learning and links to empathy. *Proc. Natl. Acad. Sci. U. S. A.* 113, 9763–9768
23. Lockwood, P.L. *et al.* (2018) Neural mechanisms for learning self and other ownership. *Nat. Commun.* 9, 4747
24. Apps, M.A.J. *et al.* (2015) Vicarious reinforcement learning signals when instructing others. *J. Neurosci.* 35, 2904–2913
25. Sul, S. *et al.* (2015) Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proc. Natl. Acad. Sci. U. S. A.* 112, 7851–7856
26. Gabay, A.S. and Apps, M.A.J. (2020) Foraging optimally in social neuroscience: computations and methodological considerations. *Soc. Cogn. Affect. Neurosci.* Published online March 30, 2020. https://doi.org/10.1093/scan/nsaa037
27. Reader, S.M. and Laland, K.N. (2002) Social intelligence, innovation, and enhanced brain size in primates. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4436
28. Dunbar, R.I.M. (1992) Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* 22, 469–493
29. Dunbar, R.I.M. (1998) The social brain hypothesis. *Evol. Anthropol. Issues News Rev.* 6, 178–190
30. Sawaguchi, T. and Kudo, H. (1990) Neocortical development and social structure in primates. *Primates* 31, 283–289
31. Jolly, A. (1985) The evolution of primate behavior: a survey of the primate order traces the progressive development of intelligence as a way of life. *Am. Sci.* 73, 230–239
32. Humphrey, N.K. (1976) The social function of intellect. In *Growing Points in Ethology* (Bateson, P.P.G. and Hinde, R.A., eds), pp. 303–318, Cambridge University Press
33. Matsuo, T. *et al.* (2015) Genetic dissection of pheromone processing reveals main olfactory system-mediated social behaviors in mice. *Proc. Natl. Acad. Sci. U. S. A.* 112, E311–E320
34. Asaba, A. *et al.* (2014) Sexual attractiveness of male chemicals and vocalizations in mice. *Front. Neurosci.* 8, 231
35. Horschler, D.J. *et al.* (2020) Do non-human primates really represent others' beliefs? *Trends Cogn. Sci.* Published online June 24, 2020. https://doi.org/10.1016/j.tics.2020.05.009
36. Grimm, C. *et al.* (2020) Shedding light on social reward circuitry: (un)common blueprints in humans and rodents. *Neuroscientist* Published online June 6, 2020. https://doi.org/10.1177/1073858420923552
37. Bartal, I.B.-A. *et al.* (2011) Empathy and pro-social behavior in rats. *Science* 334, 1427–1430
38. Laubach, M. *et al.* (2018) What, if anything, is rodent prefrontal cortex? *eNeuro* 5, e0315-18.2018
39. Joiner, J. *et al.* (2017) Social learning through prediction error in the brain. *Npj Sci. Learn.* 2, 8
40. Lee, D. and Seo, H. (2016) Neural basis of strategic decision making. *Trends Neurosci.* 39, 40–48
41. van Heukelum, S. *et al.* (2020) Where is cingulate cortex? A cross-species view. *Trends Neurosci.* 43, 285–288
42. Lockwood, P.L. (2016) The anatomy of empathy: vicarious experience and disorders of social cognition. *Behav. Brain Res.* 311, 255–266
43. Apps, M.A.J. *et al.* (2013) The role of the midcingulate cortex in monitoring others' decisions. *Front. Neurosci.* 7, 251
44. Burgos-Robles, A. *et al.* (2019) Conserved features of anterior cingulate networks support observational learning across species. *Neurosci. Biobehav. Rev.* 107, 215–228
45. Chang, S.W.C. *et al.* (2013) Neuronal reference frames for social decisions in primate frontal cortex. *Nat. Neurosci.* 16, 243–250
46. Kendal, R.L. *et al.* (2018) Social learning strategies: bridge-building between fields. *Trends Cogn. Sci.* 22, 651–665
47. Rudebeck, P.H. *et al.* (2006) A role for the macaque anterior cingulate gyrus in social valuation. *Science* 313, 1310–1312
48. Basile, B. *et al.* (2020) The anterior cingulate cortex is necessary for forming prosocial preferences from vicarious reinforcement in monkeys. *PLoS Biol.* 18, e3000677
49. Lockwood, P.L. *et al.* (2020) Anterior cingulate cortex: a brain system necessary for learning to reward others? *PLoS Biol.* 18, e3000735
50. Carrillo, M. *et al.* (2019) Emotional mirror neurons in the rat's anterior cingulate cortex. *Curr. Biol.* 29, 1301–1312
51. Hernandez-Lallement, J. *et al.* (2020) Harm to others acts as a negative reinforcer in rats. *Curr. Biol.* 30, 949–961
52. Allsop, S.A. *et al.* (2018) Corticoamygdala transfer of socially derived information gates observational learning. *Cell* 173, 1329–1342
53. Yizhar, O. *et al.* (2011) Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* 477, 171–178
54. Rudebeck, P.H. *et al.* (2007) Distinct contributions of frontal areas to emotion and social behaviour in the rat. *Eur. J. Neurosci.* 26, 2315–2326
55. Jeon, D. *et al.* (2010) Observational fear learning involves affective pain system and Cav1.2 Ca^{2+} channels in ACC. *Nat. Neurosci.* 13, 482–488
56. Niv, Y. and Langdon, A. (2016) Reinforcement learning with Marr. *Curr. Opin. Behav. Sci.* 11, 67–73
57. Schultz, W. (2013) Updating dopamine reward signals. *Curr. Opin. Neurobiol.* 23, 229–238
58. Dayan, P. and Daw, N.D. (2008) Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* 8, 429–453
59. Lockwood, P.L. *et al.* (2015) Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *J. Neurosci.* 35, 13720–13727
60. Hill, M.R. *et al.* (2016) Observational learning computations in neurons of the human anterior cingulate cortex. *Nat. Commun.* 7, 12722
61. Burke, C.J. *et al.* (2010) Neural mechanisms of observational learning. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14431–14436
62. Hackel, L.M. *et al.* (2015) Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nat. Neurosci.* 18, 1233–1235
63. Dal Monte, O. *et al.* (2020) Specialized medial prefrontal–amygdala coordination in other-regarding decision preference. *Nat. Neurosci.* 23, 565–574
64. Felix-Ortiz, A.C. *et al.* (2016) Bidirectional modulation of anxiety-related and social behaviors by amygdala projections to the medial prefrontal cortex. *Neuroscience* 321, 197–209
65. Balsters, J.H. *et al.* (2016) Disrupted prediction errors index social deficits in autism spectrum disorder. *Brain* 140, 235–246
66. Devaine, M. *et al.* (2017) Reading wild minds: a computational assay of theory of mind sophistication across seven primate species. *PLoS Comput. Biol.* 13, e1005833
67. Mitchell, J.P. (2006) Mentalizing and Marr: an information processing approach to the study of social cognition. *Brain Res.* 1079, 66–75
68. Koster-Hale, J. and Saxe, R. (2013) Theory of mind: a neural prediction problem. *Neuron* 79, 836–848
69. Apperly, I.A. (2012) What is 'theory of mind'? Concepts, cognitive processes and individual differences: *Q. J. Exp. Psychol.* 65, 825–839
70. Yoshida, W. *et al.* (2008) Game theory of mind. *PLoS Comput. Biol.* 4, e1000254
71. Hill, C.A. *et al.* (2017) A causal account of the brain network computations underlying strategic social behavior. *Nat. Neurosci.* 20, 1142–1149
72. Hampton, A.N. *et al.* (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci. U. S. A.* 105, 6741–6746
73. Devaine, M. *et al.* (2014) The social Bayesian brain: does mentalizing make a difference when we learn? *Biol.* 10, e1003992

74. Frith, C.D. and Frith, U. (2006) The neural basis of mentalizing. *Neuron* 50, 531–534
75. Hayashi, T. *et al.* (2020) Macaques exhibit implicit gaze bias anticipating others' false-belief-driven actions via medial prefrontal cortex. *Cell Rep.* 30, 4433–4444
76. Frith, C.D. (2012) The role of metacognition in human social interactions. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 2213–2223
77. Heyes, C. *et al.* (2020) Knowing ourselves together: the cultural origins of metacognition. *Trends Cogn. Sci.* 24, 349–362
78. Rilling, J.K. *et al.* (2004) The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* 22, 1694–1703
79. Gallagher, H.L. and Frith, C.D. (2003) Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* 7, 77–83
80. Apps, M.A.J. *et al.* (2013) Reinforcement learning signals in the anterior cingulate cortex code for others' false beliefs. *NeuroImage* 64, 1–9
81. Woo, C.-W. *et al.* (2014) Separate neural representations for physical pain and social rejection. *Nat. Commun.* 5, 5380
82. Corradi-Dell'Acqua, C. *et al.* (2014) Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Soc. Cogn. Affect. Neurosci.* 9, 1175–1184
83. Corradi-Dell'Acqua, C. *et al.* (2016) Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nat. Commun.* 7, 10904
84. Piva, M. *et al.* (2019) The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *eLife* 8, e44939
85. Thornton, M.A. *et al.* (2019) People represent their own mental states more distinctly than those of others. *Nat. Commun.* 10, 2117
86. Love, B.C. (2015) The algorithmic level is the bridge between computation and brain. *Top. Cogn. Sci.* 7, 230–242
87. Hillman, K.L. and Bilkey, D.K. (2012) Neural encoding of competitive effort in the anterior cingulate cortex. *Nat. Neurosci.* 15, 1290–1297
88. Hosokawa, T. and Watanabe, M. (2012) Prefrontal neurons represent winning and losing during competitive video shooting games between monkeys. *J. Neurosci.* 32, 7662–7671
89. Wittmann, M. *et al.* (2016) Self–other mergence in the frontal cortex during cooperation and competition. *Neuron* 91, 482–493
90. Meffert, H. *et al.* (2013) Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain* 136, 2550–2562
91. Popper, K.R. (1959) *The logic of scientific discovery*, Basic Books
92. Castelli, F. *et al.* (2000) Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12, 314–325
93. Stanley, J. *et al.* (2007) Effects of agency on movement interference during observation of a moving dot stimulus. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 915–926