Is there a 'social' brain? Implementations and algorithms

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

[1]Department of Experimental Psychology, University of Oxford, UK

[2]Wellcome Centre for Integrative Neuroimaging, University of Oxford, UK

[3]Centre for Human Brain Health, School of Psychology, University of Birmingham

[4]Department of Psychology, Yale University, USA

[5]Department of Neuroscience, Yale University School of Medicine, USA

[6]Kavli Institute for Neuroscience, Yale University School of Medicine, USA

Abstract

A fundamental question in psychology and neuroscience is the extent to which cognitive and neural processes are specialized for social behaviour, or 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 whether information processing is 'social' or 'non-social' can be understood 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.

Keywords

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

Correspondence: Patricia L. Lockwood, Department of Experimental Psychology, Tinsley Building, University of Oxford, OX1 3SR, United Kingdom, patricia.lockwood@psy.ox.ac.uk.
^These authors contributed equally

social situations in a way that is somehow different to what is required for 'non-social' cognitive, motor and perceptual abilities?

Here we draw on the pioneering idea of **Marr's levels** [5] to provide a framework to understand and test whether cognitive and neural processes are socially specialised, or not. We argue that a process could 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 co-operation 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 cognitive or neural processes being specialised, calling 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 in order to understand an information processing system, it is critical to consider multiple levels of explanation – the 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 algorithmic, or '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.

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 that the social process is realised in. Although the number of levels and their independence is debated [6,7], Marr's theory provides an important organising framework that suggests that social specificity can be delineated at different levels. The most critical 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 level, 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 being encoded based on an overlap in implementation (the same neuron fires similarly or the same fMRI BOLD response) to first-person and third-person events. Examples include pain to self and other [8,9], monetary and **social reward** [10,11] or one's own or other's action goals [12,13]. This overlap is interpreted as a 'common

coding' of different processes, namely that both first-person understanding and third-person understanding, or empathy has occurred [6]. However, such conclusions about mirroring are being drawn with reference only to the implementational level. But what if a different algorithm was being used to 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 the 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, so 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 in order 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? Contrasting with the algorithmic level, there is an increasing consensus for a 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 just exemplars but highlight how debates in social neuroscience and psychology can be addressed by considering which of Marr's levels one's debate surrounds. Such an idea could inform experimental design in studies of social cognition and its neural basis. Key to addressing such debates empirically will be using experimental designs in which one of the levels is held constant, either the algorithm or implementation, to test for specificity at the other. For example, lesion and brain stimulation approaches can clearly examine the impact of disrupting 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 process, and then test whether it is differentially implemented in a social (learning about rewards for others) vs 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 subsequent 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 is not necessarily given that neural systems and processes must be specialised.

Moreover, whilst 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. As well as 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, there are parallels that can be drawn, perhaps most readily in the domain of appetitive and aversive processing as 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 level?

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 as 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 sub-regions within 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 foregone reward outcomes of self. By contrast, neurons in the ACCg predominantly encoded the received reward outcome of a conspecific monkey, with some neurons responding to others' rewards exclusively (other-referenced), and another cluster showing a response that was 'mirror-like', encoding rewards of self and other (both-referenced). These results suggest at the level of single cells there is some social specificity, as there were more cells that responded to others' rewards in ACCg than in the other areas tested.

Although there were still cells that responded to the monkeys' own reward in ACCg, lesion studies can provide causal evidence as to whether a region's function disrupts social behaviour. Strikingly, when the whole ACCg region is disrupted, it is attention to social information that 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 a specific disruption to learning which stimuli rewarded others, but not self [48]. In contrast, preferences to reward self and other over neither with previously learned stimuli were preserved. It remains to be tested if these effects were largely driven by the lesion to ACCg [49]. However, they provide clear evidence that damage at the implementational level, to 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, with this effect 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], there are a larger proportion of neurons in rodent ACC that 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 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 implementation (Figure 3); there are cells in 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 on to functional anatomy [56–58]. **Reinforcement learning** (RL), is perhaps the best exemplar of a clear algorithmic process, which 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 when 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 understanding and trust learning and non-social learning about rewards are 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. In contrast, the evidence from the ACCg points to the possibility that brain areas can be socially specialised but may do without needing 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 just in single cells or brain area, but 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 BLA are specific for observational as compared to classical fear learning [52]. Activation of a basolateral amygdala➜mPFC pathway increases anxiety-like behaviours and reduces social interaction, whereas inhibiting the same pathway increases social interaction and reduced anxiety [64]. This shows that although research suggests a lack of social specificity when considering the whole amygdala [11], there could still be social specific circuit-level implementations when interacting with medial prefrontal areas. In humans, there are fewer studies that have directly compared social vs. non-social connectivity but notable examples include that ACCg-rostral cingulate connectivity 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 TPJ during social compared to direct fear learning [20]. These are just 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. But, are there algorithms that are social specific? This question is more challenging in terms of clearly defining algorithms, and there is wide controversy at 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 used 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 of 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 was one that 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 just 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 shown 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 reinforcement learning, 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 partner rodents [50,51]. There is also evidence in humans that simply forming associations between self, close others, and distant others is tracked in 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 involved in that process.

Are the algorithms of theory of mind implemented in a socially specific manner? Several studies have suggested that the temporo-parietal junction (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 problem 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 are needed 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 a distinct implementation – i.e. 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 inherent within the RSA approach is the testing of 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, 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 reinforcement learning, 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 vs. houses. Future research may be able to link levels of social specificity by using multivariate and model-based approaches. These approaches can also shed light on the nature of the levels themselves, as well as their dependence, 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. But, 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 – 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 prefrontal cortical 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, supporting the ability of social goals to regulate specificity at the other levels, has 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 algorithms used between the two groups might look like algorithmic or implementational differences, when actually it is the goal that is different between groups and causes the changes in neural response. In support of this, the neural responses of individuals with psychopathy and non-psychopathic participants were compared whilst they viewed video clips of emotional hand interactions [90]. The authors found group differences in neural responses were markedly reduced when the psychopathic offenders were instructed to empathise vs. 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. Here, we outline 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 reinforcement learning and economic decision-making, and multivariate techniques such as representational similarity analysis, will be important 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

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 fires similarly for 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 it humans.

Reinforcement learning

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 people.

References

1. Crespi BJ (2001) The evolution of social behavior in microorganisms. Trends Ecol. Evol. 16, 178–183 [PubMed: 11245940]

2. Chen P and Hong W (2018) Neural Circuit Mechanisms of Social Behavior. Neuron 98, 16–30 [PubMed: 29621486]

3. Bshary R et al. (2014) Social cognition in fishes. Trends Cogn. Sci. 18, 465–471 [PubMed: 24815200]

4. Wittmann MK et al. (2018) Neural Mechanisms of Social Cognition in Primates. Annu. Rev. Neurosci. DOI: 10.1146/annurev-neuro-080317-061450

5. Marr D (1982) Vison.

6. Krakauer JW et al. (2017) Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93, 480–490 [PubMed: 28182904]

7. Bickle J (2015) Marr and Reductionism. Top. Cogn. Sci. 7, 299–311 [PubMed: 25772159]

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 [PubMed: 20946964]

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 [PubMed: 20974173]

10. Ruff CC and Fehr E (2014) The neurobiology of rewards and values in social decision making. Nat. Rev. Neurosci. 15, 549–562 [PubMed: 24986556]

11. Munuera J et al. (2018) Shared neural coding for social hierarchy and reward value in primate amygdala. Nat. Neurosci. 21, 415–423 [PubMed: 29459764]

12. Umilta MA et al. (2001) I know what you are doing: A neurophysiological study. Neuron 31, 155–165 [PubMed: 11498058]

13. Chong TT-J et al. (2008) fMRI adaptation reveals mirror neurons in human inferior parietal cortex. Curr. Biol. 18, 1576–80 [PubMed: 18948009]

14. Heyes C (2012) What's social about social learning? J. Comp. Psychol. 126, 193–202 [PubMed: 21895355]

15. Cook R et al. (2014) Mirror neurons: from origin to function. Behav. Brain Sci. 37, 177–192 [PubMed: 24775147]

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 TEJ et al. (2008) Associative learning of social value. Nature 456, 245–249 [PubMed: 19005555]

18. Lockwood PL and Klein-Flügge M Computational modelling of social cognition and behaviour—a reinforcement learning primer. Soc. Cogn. Affect. Neurosci. DOI: 10.1093/scan/nsaa040

19. Olsson A et al. (2020) The neural and computational systems of social learning. Nat. Rev. Neurosci. DOI: 10.1038/s41583-020-0276-4

20. Lindström B et al. (2018) A common neural network differentially mediates direct and social fear learning. NeuroImage 167, 121–129 [PubMed: 29170069]

21. Apps MAJ et al. (2016) The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. Neuron 90, 692–707 [PubMed: 27196973]

22. Lockwood PL et al. (2016) Neurocomputational mechanisms of prosocial learning and links to empathy. Proc. Natl. Acad. Sci. 113, 9763–9768 [PubMed: 27528669]

23. Lockwood PL et al. (2018) Neural mechanisms for learning self and other ownership. Nat. Commun. 9, 4747 [PubMed: 30420714]

24. Apps MAJ et al. (2015) Vicarious reinforcement learning signals when instructing others. J. Neurosci. 35, 2904–2913 [PubMed: 25698730]

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. 112, 7851–7856 [PubMed: 26056280]

26. Gabay AS and Apps MAJ. Foraging Optimally in Social Neuroscience: Computations and Methodological considerations. Soc. Cogn. Affect. Neurosci. DOI: 10.1093/scan/nsaa037

27. Reader SM and Laland KN (2002) Social intelligence, innovation, and enhanced brain size in primates. Proc. Natl. Acad. Sci. 99, 4436 [PubMed: 11891325]

28. Dunbar RIM (1992) Neocortex size as a constraint on group size in primates. J. Hum. Evol. 22, 469–493

29. Dunbar RIM (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 NK (1976) The social function of intellect In Growing points in ethology Cambridge U 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. 112, E311–E320 [PubMed: 25564662]

34. Asaba A et al. (2014) Sexual attractiveness of male chemicals and vocalizations in mice. Front. Neurosci. 8, [PubMed: 24550770]

35. Horschler DJ et al. (2020) Do Non-Human Primates Really Represent Others' Beliefs? Trends Cogn. Sci 0,

36. Grimm C et al. (2020) Shedding Light on Social Reward Circuitry: (Un)common Blueprints in Humans and Rodents. The Neuroscientist DOI: 10.1177/1073858420923552

37. Bartal IB-A et al. (2011) Empathy and Pro-Social Behavior in Rats. Science 334, 1427–1430 [PubMed: 22158823]

38. Laubach M et al. (2018) What, If Anything, Is Rodent Prefrontal Cortex? eNeuro 5,

39. Joiner J et al. (2017) Social learning through prediction error in the brain. Npj Sci. Learn. 2, 8 [PubMed: 30631454]

40. Lee D and Seo H (2016) Neural Basis of Strategic Decision Making. Trends Neurosci. 39, 40–48 [PubMed: 26688301]

41. Heukelum S van et al. (2020) Where is Cingulate Cortex? A Cross-Species View. Trends Neurosci. 0,

42. Lockwood PL (2016) The anatomy of empathy: Vicarious experience and disorders of social cognition. Behav. Brain Res. 311, 255–266 [PubMed: 27235714]

43. Apps MAJ et al. (2013) The role of the midcingulate cortex in monitoring others' decisions. Front. Neurosci. 7, 251 [PubMed: 24391534]

44. Burgos-Robles A et al. (2019) Conserved features of anterior cingulate networks support observational learning across species. Neurosci. Biobehav. Rev. 107, 215–228 [PubMed: 31509768]

45. Chang SWC et al. (2013) Neuronal reference frames for social decisions in primate frontal cortex. Nat. Neurosci. 16, 243–250 [PubMed: 23263442]

46. Kendal RL et al. (2018) Social Learning Strategies: Bridge-Building between Fields. Trends Cogn. Sci. 22, 651–665 [PubMed: 29759889]

47. Rudebeck PH et al. (2006) A role for the macaque anterior cingulate gyrus in social valuation. Science 313, 1310–1312 [PubMed: 16946075]

48. Basile B et al. (in press) The anterior cingulate cortex is necessary for forming prosocial preferences from vicarious reinforcement in monkeys. PLOS Biol.

49. Lockwood PL et al. (in press) Anterior cingulate cortex: A brain system necessary for learning to reward others? PLOS Biol.

50. Carrillo M et al. (2019) Emotional Mirror Neurons in the Rat's Anterior Cingulate Cortex. Curr. Biol. 29, 1301–1312.e6 [PubMed: 30982647]

51. Hernandez-Lallement J et al. (2020) Harm to Others Acts as a Negative Reinforcer in Rats. Curr. Biol. 30, 949–961.e7 [PubMed: 32142701]

52. Allsop SA et al. (2018) Corticoamygdala Transfer of Socially Derived Information Gates Observational Learning. Cell 173, 1329–1342.e18 [PubMed: 29731170]

53. Yizhar O et al. (2011) Neocortical excitation/inhibition balance in information processing and social dysfunction. Nature 477, 171–178 [PubMed: 21796121]

54. Rudebeck PH et al. (2007) Distinct contributions of frontal areas to emotion and social behaviour in the rat. Eur. J. Neurosci. 26, 2315–26 [PubMed: 17927774]

55. Jeon D et al. (2010) Observational fear learning involves affective pain system and Cav1.2 Ca2+ channels in ACC. Nat. Neurosci. 13, 482–8 [PubMed: 20190743]

56. Niv Y and Langdon A (2016) Reinforcement learning with Marr. Curr. Opin. Behav. Sci. 11, 67–73 [PubMed: 27408906]

57. Schultz W (2013) Updating dopamine reward signals. Curr. Opin. Neurobiol. 23, 229–238 [PubMed: 23267662]

58. Dayan P and Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. 8, 429–453 [PubMed: 19033240]

59. Lockwood PL et al. (2015) Encoding of Vicarious Reward Prediction in Anterior Cingulate Cortex and Relationship with Trait Empathy. J. Neurosci. 35, 13720–13727 [PubMed: 26446224]

60. Hill MR et al. (2016) Observational learning computations in neurons of the human anterior cingulate cortex. Nat. Commun. 7, 1–12

61. Burke CJ et al. (2010) Neural mechanisms of observational learning. Proc. Natl. Acad. Sci. U. S. A. 107, 14431–14436 [PubMed: 20660717]

62. Hackel LM et al. (2015) Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. Nat. Neurosci. 18, 1233–1235 [PubMed: 26237363]

63. Dal Monte O et al. (2020) Specialized medial prefrontal–amygdala coordination in other-regarding decision preference. Nat. Neurosci. 23, 565–574 [PubMed: 32094970]

64. Felix-Ortiz AC et al. (2016) Bidirectional modulation of anxiety-related and social behaviors by amygdala projections to the medial prefrontal cortex. Neuroscience 321, 197–209 [PubMed: 26204817]

65. Balsters JH 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 [PubMed: 29112973]

67. Mitchell JP (2006) Mentalizing and Marr: An information processing approach to the study of social cognition. Brain Res. 1079, 66–75 [PubMed: 16473339]

68. Koster-Hale J and Saxe R (2013) Theory of mind: a neural prediction problem. Neuron 79, 836–848 [PubMed: 24012000]

69. Apperly IA (2012) What is "theory of mind"? Concepts, cognitive processes and individual differences: Q. J. Exp. Psychol. at <https://journals.sagepub.com/doi/10.1080/17470218.2012.676055>

70. Yoshida W et al. (2008) Game Theory of Mind. Plos Comput. Biol 4,

71. Hill CA et al. (2017) A causal account of the brain network computations underlying strategic social behavior. Nat. Neurosci. 20, 1142–1149 [PubMed: 28692061]

72. Hampton AN et al. (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. U. S. A. 105, 6741–6746 [PubMed: 18427116]

73. Devaine M et al. (2014) The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? PLoS Comput. Biol 10, e1003992 [PubMed: 25474637]

74. Frith CD and Frith U (2006) The neural basis of mentalizing. Neuron 50, 531–534 [PubMed: 16701204]

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.e5 [PubMed: 32234478]

76. Frith CD (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 [PubMed: 32298621]

78. Rilling JK et al. (2004) The neural correlates of theory of mind within interpersonal interactions. Neuroimage 22, 1694–1703 [PubMed: 15275925]

79. Gallagher HL and Frith CD (2003) Functional imaging of 'theory of mind.' Trends Cogn. Sci. 7, 77–83 [PubMed: 12584026]

80. Apps MAJ et al. (2013) Reinforcement learning signals in the anterior cingulate cortex code for others' false beliefs. NeuroImage 64, 1–9 [PubMed: 22982355]

81. Woo C-W et al. (2014) Separate neural representations for physical pain and social rejection. Nat. Commun. 5, 5380 [PubMed: 25400102]

82. Corradi-Dell'acqua C et al. (2013) 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. 4, [PubMed: 22446298]

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, 1–12

84. Piva M et al. (2019) The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. eLife 8, e44939 [PubMed: 31192786]

85. Thornton MA et al. (2019) People represent their own mental states more distinctly than those of others. Nat. Commun. 10, 1–9 [PubMed: 30602773]

86. Love BC (2015) The Algorithmic Level Is the Bridge Between Computation and Brain. Top. Cogn. Sci. 7, 230–242 [PubMed: 25823496]

87. Hillman KL and Bilkey DK (2012) Neural encoding of competitive effort in the anterior cingulate cortex. Nat. Neurosci. 15, 1290–1297 [PubMed: 22885851]

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 [PubMed: 22649245]

89. Wittmann M et al. (2016) Self-Other Mergence in the Frontal Cortex during Cooperation and Competition. Neuron 91, 482–493 [PubMed: 27477020]

90. Meffert H et al. (2013) Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. Brain 136, 2550–2562 [PubMed: 23884812]

91. Popper K (2005) The logic of scientific discovery, Routledge.

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 [PubMed: 10944414]

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 [PubMed: 17683237]

Box 1: The importance of non-social control conditions

When designing an experiment to test if 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 if 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 or their conspecific [45,63]. In rodent studies of observational fear conditioning, a more typical classical conditioning condition, without any social context, has been used 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 used to try to show specificity to 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 being 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, which 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 vs. 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 pre-recorded 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 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 vs. non-social is critical for creating social and non—social experimental conditions.

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 resolution of methods are some of the most critical. Despite these challenges, comparisons can be drawn and are crucial 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. On the flip side, 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 key role of the anterior cingulate gyrus (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 suggesting social specificity at the level of algorithm and implementation is an approach that can readily be applied in different species and may allow us to draw greater connections between findings at multiple levels in future research.

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. Existing evidence in 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 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 social specificity of a cognitive or neural process? We suggest the importance of holding a level (computational, algorithm or implementation) constant while examining the impact on the other level.

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

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

Highlights

- A central question in psychology and neuroscience is the extent to which social behaviour is sub-served 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 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 anterior cingulate gyrus as crucial for processing social implementations and 'theory of mind' as a putative social algorithm.
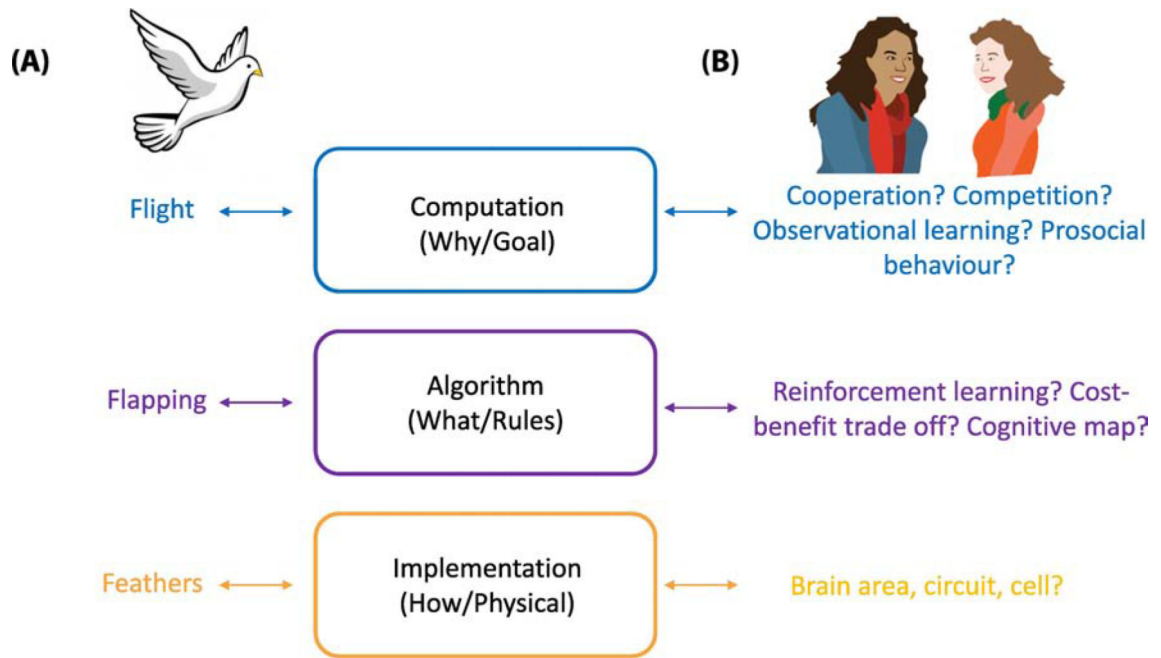
Figure 1. Marr's three levels of analysis for non-social and social behaviour.
(A) If we want to understand the goal of a bird to fly, we cannot simply study its feathers. We need to know that the bird's goal is to fly (computation), which it does by flapping its wings (algorithm) where the aerodynamics to fly depe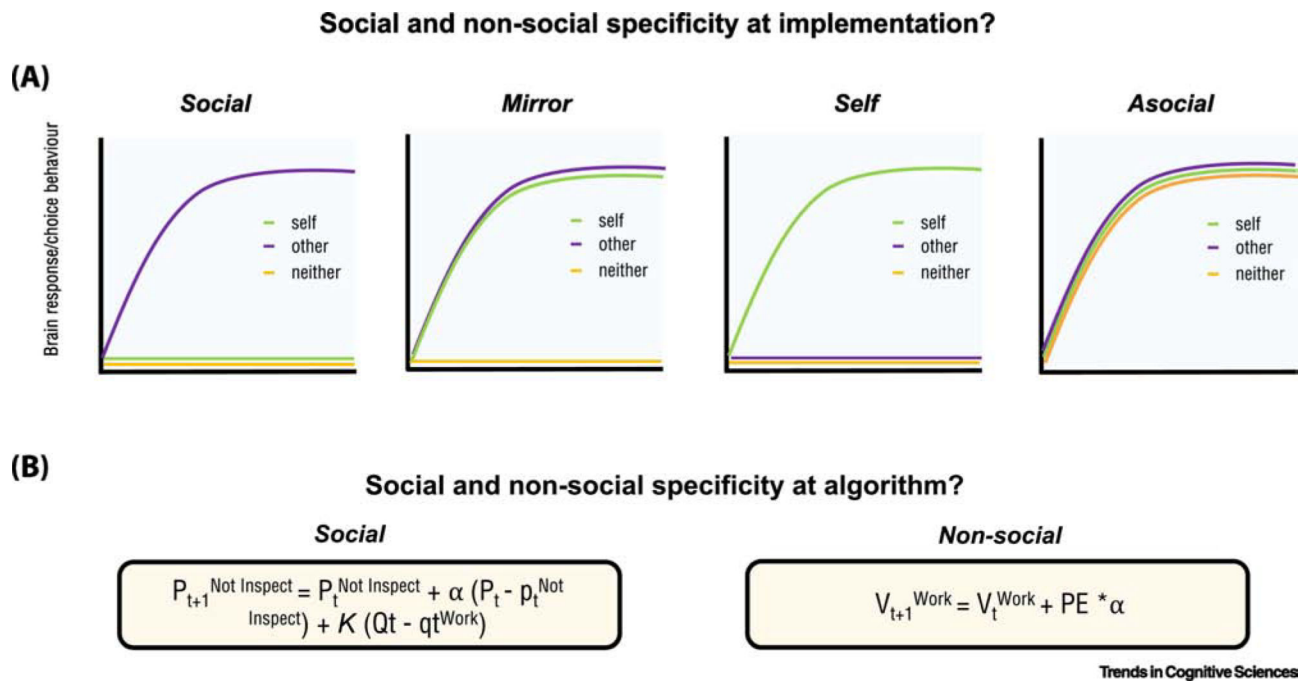nd on the feathers (implementation). (B) We argue that to understand how specific social behaviour is, as compared to other 'non-social' processes, such as bird flight, we need to understand the social goal (are we cooperating, learning from or helping the other person or group). Next, we need to understand 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, which brain areas, circuit or cell is it realized in? Critically, 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.
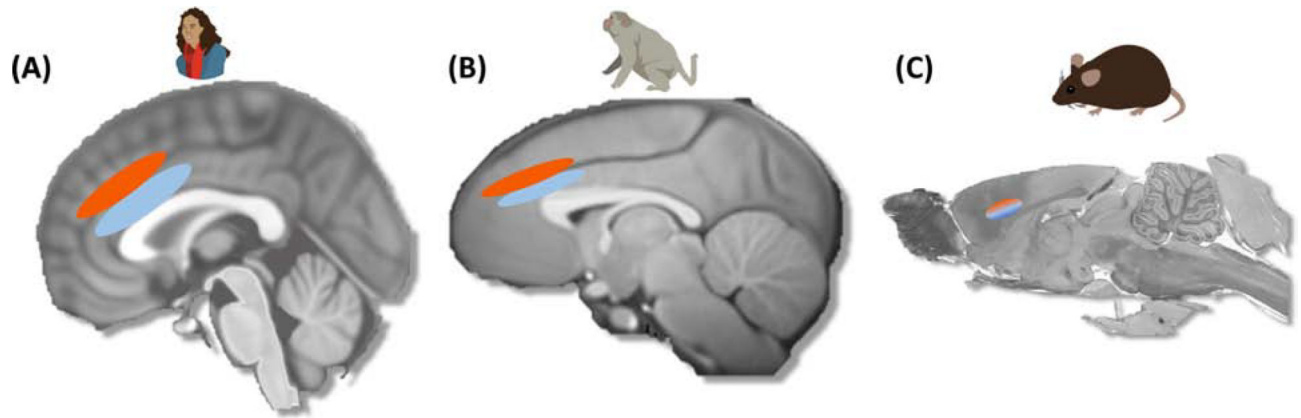
Figure 2. The anterior cingulate cortex and socially specific implementation.
(A) In humans, several studies have found that the gyral portion of the anterior cingulate cortex (ACCg, light blue) responds to prediction errors and rewards predicted or delivered to other people, whereas the neighboring sulcus (ACCs, red) responds to self-relevant reward signals and prediction errors [4,17,21,42,43]. Importantly, there is also evidence that anterior cingulate sulcus responds to forgone prediction errors and rewards hinting that this area might process a domain general response to rewards not delivered to oneself, rather than a socially specific signal [4,17,21,42,43]. (B) Converging evidence from focal lesion and single-unit recording studies suggests that ACCg responds to other monkeys' rewards in terms of attention allocation and behavioural choice, where monkeys show a preference to reward others over neither. Instead in ACCs, a large proportion of neurons signal both self reward and neither reward, consistent with an involvement in forgone rewards [45,47,48]. (C) In rodents, clear divisions of the sulcus and gyrus as not as readily apparent as they are in humans and monkeys, but roughly correspond to what is labelled 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 as to 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.

Social and non-social specificity at implementation?

(A)

(B)

Social and non-social specificity at algorithm?

Social

$$P_{t+1}^{\text{Not Inspect}} = P_t^{\text{Not Inspect}} + \alpha \left(P_t - p_t^{\text{Not Inspect}}\right) + K \left(Q_t - q_t^{\text{Work}}\right)$$

Non-social

$$V_{t+1}^{\text{Work}} = V_t^{\text{Work}} + PE * \alpha$$

Trends in Cognitive Sciences

Figure 3. Social specificity at implementation and algorithm

(A) Schematic illustrating hypothesized 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) was not able to 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 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 employers inferred probability that the employee will work. In the non-social example, Vt is 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].