Accepted for publication by *Synthese,* 20th March 2024, to appear in a topical collection on ‘Scientific Realism in Cognitive Neuroscience’, edited by Inês Hipólito, Thomas van Es, Guilherme Sanches Oliveira, and Matteo Colombo.

SOCIAL LEARNING IN MODELS AND MINDS

Daniel Yon[1] & Cecilia Heyes[2]

[1]Department of Psychological Sciences

Birkbeck, University of London, London, UK

d.yon@bbk.ac.uk

orcid.org/0000-0002-5511-3884

[2]All Souls College and Department of Experimental Psychology

University of Oxford, Oxford, UK

Corresponding author: Cecilia.heyes@all-souls.ox.ac.uk

orcid.org/0000-0001-9119-9913

SOCIAL LEARNING IN MODELS AND MINDS

Abstract

After more than a century in which social learning was blackboxed by evolutionary biologists, psychologists and economists, there is now a thriving industry in cognitive neuroscience producing computational models of learning from and about other agents. This is a hugely positive development. The tools of computational cognitive neuroscience are rigorous and precise. They have the potential to prise open the black box. However, we argue that, from the perspective of a scientific realist, these tools are not yet being applied in an optimal way. To fulfil their potential, the shiny new methods of cognitive neuroscience need to be better coordinated with old-fashioned, contrastive experimental designs. Inferences from model complexity to cognitive complexity, of the kind made by those who favour lean interpretations of behaviour ('associationists'), require social learning to be tested in challenging task environments. Inferences from cognitive complexity to social specificity, made by those who favour rich interpretations ('mentalists'), call for non-social control experiments. A parsimonious model that fits current data is a good start, but carefully designed experiments are needed to distinguish models that tell us how social learning *could* be done from those that tell us how it is *really* done.

Keywords: cognitive neuroscience; contrastive testing; domain-specificity; model complexity; scientific realism; social learning

Social learning is everywhere – in the animal kingdom, in human lives, and now in computational cognitive neuroscience. Social learning happens whenever one agent learns something from or about another agent. Snails do it when they follow the slime trails of other snails to food, and you, dear reader, are doing it right now as you learn about snails and trails from the words on this page.

For decades, the motley, bulging sack of social learning was stuffed with examples and carried to explanatory tasks by scientists interested in the consequences of learning from others. Behavioural ecologists, ethologists, comparative psychologists, and cultural evolutionists asked how social learning contributes to behavioural adaptation – how learning from others enables the behaviour of agents to meet and change the demands of their environments. Similarly, but with a focus on human social learning, developmental and educational psychologists asked how learning from others contributes to the growth of technical and social skills – learning to cook, speak, and defer to elders – while social psychologists and economists studied the roles of social learning in generating political, market, and consumer behaviour. These groups of highly productive scientists were and are concerned about evolutionary, developmental, and societal processes that shape and are shaped by social learning. But, until recently, social learning was a black box. Almost no one was asking about the psychological and neurobiological processes of social learning – about what goes on between an agent's ears when they learn from another agent (Heyes 1994; 2016; 2018).

Now, there is a thriving industry in cognitive neuroscience asking exactly this question and using sophisticated computational and neurobiological methods to tackle it. We regard this as a hugely positive development. The tools of computational cognitive neuroscience are rigorous and precise. They have the potential to prise open the black box. However, we argue that, from the perspective of a scientific realist – a scientist who wants their work to yield true, or as-true-as-possible, descriptions of cognitive processes - these tools are not yet being applied in an optimal way. To fulfil their potential, the shiny new methods of cognitive neuroscience need to be better coordinated with old-fashioned experimental design. To be clear, experiments have not been edged out by modelling in this field. Animated by a spirit of methodological pluralism (Heesen, Bright & Zucker 2019), and naturally given that cognitive neuroscience is rooted in experimental psychology, most articles in the cognitive neuroscience of social learning report experimental work as well as their models. However, we will argue that these experiments are not of the right kinds to test the researchers' hypotheses about the mechanisms of social learning.

We begin with some background on the questions asked and methods used by scientists interested in social learning. We then suggest that recent work, using computational modelling, has two inferential problems – inferring cognitive complexity from model complexity, and social specificity from cognitive complexity. In the two sections that follow, we present case studies illustrating these problems and indicating how they can be solved by contrastive experiments. Before closing, we suggest that the inferential problems we highlight, although pressing for a scientific realist, would dissolve under an instrumentalist or historicist gaze.

Questions and methods

What is special about social learning? This question has dominated scientific work on social learning since it began more than a century ago (e.g., Thorndike 1898). In the era before cognitive neuroscience, those advocating the specialness of social learning tended to contrast it with 'individual' or 'asocial' learning – learning without assistance from other agents – and to stress social learning's distinctive potential to reduce the costs of learning for individuals; to change the selection pressures acting on a gene pool, and thereby the course of a population's evolution; to produce 'traditions', systematic variation in behaviour across populations within a species; and, as an analogue of DNA replication, to underwrite a Darwinian process of cultural selection working sometimes in opposition and sometimes in concert with genetic evolution. Enthusiasm for these well-founded hypotheses - and latterly for classical evolutionary psychology - sometimes spilled over into casual claims that social and asocial learning depend on different psychological processes; that social learning is an 'evolved module' or 'adaptive specialisation'; that minds use fundamentally different, domain-specific algorithms to get information from others. But the nature of the alternative processes was never specified; the black box of social learning remained firmly shut (Heyes & Galef 1996; Hoppitt & Laland 2013; Mobius & Rosenblat 2014; Zentall & Galef 1988).

With the advent of computational cognitive neuroscience, the question of what is special about social learning continues to dominate but it has become more focussed. Cognitive neuroscientists want to know what is special about *human* social learning – how it differs from social learning in other animals, and how those differences contribute to making human lives so peculiar compared with the lives of other species (Gweon 2021). Theories draw on evidence from neurobiological studies of rats and monkeys (e.g., Lockwood, Apps & Chang 2020; Roumazeilles et al 2021), and behavioural research with children (e.g., Gweon 2021; Ho, MacGlashan, Littman & Cushman 2017; Ho, Cushman, Littman & Austerweil 2019), but most studies involve adult human participants performing social learning tasks online or in the laboratory. In the latter case, participants are often tested in a scanner, using functional magnetic resonance imaging (fMRI). The tasks typically require people to make a simple choice (e.g., between two cards) on many successive occasions (in repeated 'trials'), with 'advice' from another agent.

In addition to human-specificity, cognitive neuroscientists are interested in social-specificity at both psychological and neurobiological levels of explanation (Lockwood et al. 2020). Does social learning, not only draw information from distinctive sources – other agents as well as the inanimate environment – but also use special kinds of representation or algorithm to process and encode that information? Which brain regions, circuits, and neurochemicals are recruited for social but not asocial learning, or more for one than the other?[1]

[1] These questions are major drivers of research but it is striking that, compared with evolutionary psychologists, cognitive neuroscientists working on social learning lean towards domain-generalism. Even those who believe that social learning involves representations with highly distinctive content (e.g., representations of the minds of other agents) – and that social information is processed in dedicated circuits (e.g., in ACC gyrus), assume that the heavy lifting of social learning is done by the same psychological processes that, for example, track causal relationships among inanimate events (Velez & Gweon 2021).

Cross-cutting questions about social-specificity, in a cryptic way, are questions about the complexity or richness of social learning. In the most prominent contemporary debate, one group – we will call them *associationists* - advances the view that human social learning often depends on the same, simple processes that mediate asocial learning in humans and other animals. Much of what we know about these processes originated in research on rats and pigeons in the behaviourist era, and in some instances modern associationists make their case by revealing similarities between humans and other animals. For example, recent work has found that precisely the same computational model can be used to capture paranoid beliefs about the social world in human volunteers and the patterns of behaviour rats show as they learn to poke their noses in different ports to earn satisfying squirts of sugar (Reed et al, 2020).

The other group – let's call them *mentalists* – challenge this view, saying that, in humans at least, the processes of social learning are rarely so 'lean'; they typically involve 'rich' representations and 'inferences' about the minds of other agents (Velez & Gweon 2021). Note that the richness debate is about what is common or typical of human social learning. Rhetoric aside, it is unlikely that any associationist would deny that social learning in adult humans *sometimes* involves what is known as mentalising (or mindreading, or theory of mind). Similarly, it is unlikely that any mentalist would deny that social learning *sometimes* – in snails if not in humans – involves associative processes of the kind found in rats pressing levers for food.

However, when defining what makes a given cognitive process 'rich' or 'lean', researchers on both sides of the debate have tended to focus on the complexity or simplicity of the computational models deployed to explain behaviour. Here, 'complexity' is to be understood in terms of the number of moving parts an agent tracks when learning about their environment. On one side of the debate, associationists suppose that social learners deploy algorithms in their heads that only track a single variable – like the history of reward. In the same way that a rat can learn through trial and error which spout will yield the most sugar just by integrating the history of rewards, social learners might simply integrate the trials and errors of other agents to determine which options are best too. This scheme doesn't require representing anything about the minds of other agents, only some equivalence between rewards generated by our own actions and those generated by others (Olsson, Knapska & Lindstrom, 2020). On the other side of the debate, the mentalists suppose that social learners track a whole host of variables – such as the intentions, character or affiliations of the agents we are learning from. The complexities of social world, it is argued, necessitate more complex algorithms that track more than simply the probability of reward (Olsson, Knapska & Lindstrom, 2020). Set up in this way, the debate between associationists and mentalists is operationalised as a debate about complexity. If it can be shown that social learning can be captured by simpler mathematical models tracking only a single variable (e.g., reward history), this is taken as evidence that social learning depends on domain-general associative processes unfolding in the learner's mind. In contrast, if social learning is better captured by more complex mathematical models tracking a larger number of variables – combining say, reward history, with the intentions and characters of our social partners – this is taken as evidence that social learning depends on specific inferential processes that are only deployed when dealing with the social world.

Model complexity > Cognitive complexity > Social specificity

Examined more closely then, computational cognitive scientists who are interested in social learning use models to find out about minds in a three-step process:

1. Model complexity. A modeller establishes whether behaviour (or neural) data from a task are best fit by a simple or complex model.
2. Cognitive complexity. The results of Step 1 licenses a conclusion that the cognitive processes engaged by the task are either simple or complex.
3. Social specificity. The result of Step 2 licenses a conclusion that these same processes are domain-general (if simple) or socially specific (if complex).

For this method to work, each step must be valid. In the associationist case, it must be valid to infer simple cognitive processes from the fit of a simple model (Step 1 > Step 2) *and* to infer domain-generality from cognitive simplicity (Step 2 >

Step 3). In the mentalist case, it must be valid to infer complex cognitive processes from the fit of a complex model (Step 1 > Step 2) *and* to infer social specificity from cognitive complexity.

In our view, each camp has a problem. The problem for the associationist relates to the first inference, from Step 1 to Step 2. Model comparison favouring a 'simple' model in two different conditions cannot rule out the possibility that different kinds of cognitive process support the same algorithm in the two conditions. For example, if the same prediction-error algorithm fits the data when participants are using direct feedback and advice, it does not necessarily mean that the 'predictions' or 'errors' were generated by lean, domain-general cognitive processes in both conditions.

The problem for the mentalist is different. Model comparison logic means that, if we have a simple model and a more complex model, and the more complex model wins, we are justified in preferring not only the more complex model but also a theory that hypothesises a more complex cognitive process; Step 1 > Step 2 is valid. The problem for the mentalist is in the second inference from Step 2 > Step 3, cognitive complexity to social specificity. Behaviour in a task may depend on a more complex cognitive process without the process being socially specific. For example, I might create a model-based algorithm, where I, the modeller, designate the state the agent is learning about as an internal state of another agent – a state, such as a belief or competence, of a kind that is only possessed by other agents. If this model then beats a simple learning algorithm, I have reason to believe the modelled behaviour depended on a more complex cognitive process, but not that my designation of the states was the correct one. The learner might have been tracking an external, observable state of another agent and/or of the context – states of a kind that characterise the inanimate world as well as other agents.

In other words, both camps have problems of contrastive underdetermination of theory by evidence (Duhem 1991; Quine 1958; Stanford 2006). In the associationist case, the success of a simple model is compatible with a theory suggesting that the same, simple psychological processes mediate social and asocial learning. But it is also compatible with a theory suggesting that different psychological processes mediate social and asocial learning. In the mentalist case, the success of a complex model is compatible with a theory suggesting that the modelled example of social learning is mediated by a complex, socially specific psychological process. But it is also compatible with a theory suggesting that the example depends on a complex, domain-general psychological process.

Scientists can respond to problems of contrastive underdetermination by invoking a general principle to justify their theory choice, or by doing further experimental work to discover which of the 'tied' theories is more empirically adequate. Taking the first option, associationists and mentalists could – and sometimes do – appeal to a principle of parsimony or simplicity. This is unsatisfactory for many reasons unearthed in debate about Morgan's Canon (Dacey 2016, 2021; Fitzpatrick 2008, 2017). For example, there are a variety of different kinds of explanatory simplicity (ontological, iterative, phylogenetic, uniformity, ease of use), they are not highly correlated, and there is no well-established epistemic basis for prioritising one kind over another. Similarly, mentalists could – but do not - argue that there is a compelling body of evidence that, as claimed by evolutionary psychologists, complex cognitive processes tend to be domain-specific rather than domain-general. Any argument of this kind is unlikely to be convincing given widespread concern about the empirical bases and conceptual coherence of evolutionary psychology's version of the modularity of mind thesis (Pietraszewski & Wertz 2021; Robbins 2013).

Given the weakness of parsimony and modularity arguments, we recommend that computational cognitive scientists working on social learning take the second option – run more experiments of specific kinds. If our diagnosis of the problems is correct, they can be remedied by backing up the computational modelling with better (and more old-fashioned) experimental manipulations. Associationists who want to secure claims about domain-general simplicity need to test their models in challenging conditions where complex cognitive processes have an opportunity to shine. Mentalists who want to secure claims about domain-specific complexity need to show their complex algorithms are only deployed in social settings. This will mean crafting non-social control conditions that avoid a social-asocial / simple-complex confound.

The associationist case

The problem for those who favour a simple, associationist view of social learning can be illustrated by a seminal study by Behrens, Hunt, Woolrich & Rushworth (2008). In this experiment, volunteers performed a simultaneously social and asocial learning task. On each trial, participants had to choose between two coloured rectangles presented on screen – one of which would tend to yield a monetary reward, and one of which would tend not to. To line their pockets as much as possible, each participant could exploit two sources of information about where the rewards would be. First, the volunteers could use their own direct experience of reward history – tracking which rectangle had yielded the most rewards in the past. Alongside this direct route (asocial), participants could also use advice provided by another agent. On each trial, this social advice from the other agent was represented by a red frame that appeared around the recommended rectangle before the participant made their decision.

Behrens and colleagues showed that participants' learning about both sources of information – direct experience and social advice – can be modelled using a Bayesian reinforcement learning algorithm. Moreover, recording the brain activity of volunteers as they learn about both types of information reveals a signature of the prediction error computations hypothesised by the algorithm. Since this kind of prediction error-based learning is at the heart of associationist accounts of reward learning in creatures like pigeons and rats, the authors conclude that social learning 'can be realised by means of the same associative processes previously established for learning other, simpler, features of the environment' (p. 245).

This is a seductive conclusion, but it might not be a sound one. This is because it depends on a kind of *reverse inference*. The terms *forward inference* and *reverse inference* have been used for over a decade in cognitive neuroscience to characterise different claims researchers can make about the relationship between brain structures and cognitive functions (Poldrack, 2006). A *forward inference* involves mapping from function-to-structure. For instance, a researcher might design a brain imaging experiment where one condition involves reasoning about mental states, and another condition does not. By contrasting brain activity in the first condition and the second, the scientist can identify which brain regions are involved in mentalising – say, the medial prefrontal cortex (mPFC). In contrast, a *reverse inference* involves mapping from structure-to-function. Here, a researcher might look at brain activity they have recorded in another experiment and notice that they have found activity in the mPFC. Since previous studies have implicated the mPFC in reasoning about mental states, they conclude that their participants were mentalising in this study too – a *reverse inference*. These reverse inferences are not deductively valid and can lead neuroscientists astray when brain regions show little selectivity in the tasks that engage them (Poldrack, 2011).

Here, we can see an analogous kind of *computational reverse inference,* with researchers mapping algorithms rather than brain regions onto hidden cognitive processes. First, the researchers establish that a particular learning algorithm can account for behaviour and brain activity in a simple reward learning task (Behrens, Woolrich, Walton & Rushworth 2007) – a *forward inference*. Then, the authors show that the same algorithm can also effectively model behaviour and brain activity during a social learning task. The success of the algorithm in the second instance is taken to imply that the same kind of simple cognitive processes are involved in the social case too – a *reverse inference*.

The issue with the shape of this inference is that rich, complex social cognitive processes could also yield patterns of behaviour and neural activity that can be successfully captured by simple algorithms.

This can be made clearer if we consider the details of the study by Behrens et al (2008) in full. At the beginning of the experiment, participants are introduced to a confederate who they believe will be providing the advice during the main test phase. The experimenter explains to them both that while the learner's job is to earn as much as possible, the confederate's task is to keep the learner's winnings within a particular unknown range – not too high or too low. This means that while the advisor always knows for certain which rectangle will be rewarded, they will sometimes be incentivised to provide accurate advice, and other times incentivised to mislead.

Now, against this backdrop, it is perfectly plausible that the associationist story is true. Participants might learn about the value of the red frame simply by tracking its reward history from trial to trial, without deploying any rich, mentalistic processes. This is the interpretation that Behrens et al (2008) favour. However, it is also possible that participants in this task *do* deploy a specialised suite of mentalising processes to solve the social learning element of the task. For example, a learner might construct a mentalistic model of the unobserved advisor, constructing rich representations of what they intend, desire and believe. For instance, such a model might attribute to the advisor a desire to limit the learner's earnings, and thus an intention to deceive with their next recommendation.

The problem with distinguishing between these two accounts using algorithms is that both the rich and the lean accounts assume agents use the same information (the red frame) to generate and update a sequence of predictions. While the prediction error signals identified by Behrens et al (2008) could arise through a simple, domain-general associative process, the same trajectory of predictions and prediction errors would be generated by a richer mentalising process that uses the red frame to infer and attribute desires and beliefs. This means, in principle, that a complex mentalising agent may generate behaviours that can be easily described by a simple learning algorithm.

A common response to this kind of theoretical dilemma is to invoke the principle of parsimony. Indeed, associationists have often suggested that simple learning processes should act as a 'null hypothesis' that psychologists need to reject before supplying richer alternatives (e.g., Haselgrove 2016; Macphail 1985). Parsimony may sometimes be a virtue, but as our sketch of Behrens et al (2008) shows, we may cut ourselves on Occam's razor if we always plump for the leanest possible account (Heyes 2012; Sober 2009). Indeed, if we are scientific realists, we want our scientific processes to pick out theories that are likely to map onto genuine reality, rather than picking out the simplest possible theory that would account for our data.

If we cannot rely on parsimony to adjudicate between competing theories, our only option is to bite the empirical bullet, and to run experiments aimed at contrastive testing[2]. In this way of thinking, we can distinguish between associationist and mentalist accounts not by appealing to the simplest possible model, but by designing experiments where the two models make competing predictions. For example, a key prediction of the mentalistic alternative we offer above is that learners construct an internal model that attributes beliefs and desires to another agent. One way of ruling out this possibility more effectively would be to include experimental conditions where these more complex mentalistic inferences would lead to predictions that diverge from the simpler algorithm.

For example, participants could be presented with a version of the task with a different regime of advice and slightly tweaked cover story. Imagine instead that participants are explicitly told that the advisor's goal is not to keep participant earnings in some unknown range, but to make recommendations that ensure the participants get half of their choices rewarded, but half unrewarded. They will achieve this by adjusting the advice they give based on how the learner is behaving trial-by-trial: if your last trial was rewarded, the advisor will be more likely to make a deceptive, incorrect recommendation – but if your last trial was unsuccessful, the advisor will be more likely to make a cooperative, correct recommendation.

In such a variant of the task, it would be possible to present participants with similar regimes of advice – but the 'prediction errors' generated by associationist and mentalist computations would dramatically diverge.

The associationist account assumes that participants are simply learning about the reward history of the red frame, and thus any prediction error associated

[2] Following Currie (2021), we can think that 'parsimony' in cognitive science is related to our current knowledge of a creature's cognitive capacities: if we know that a creature already possesses a certain cognitive capacity, and this capacity is enough to explain the behaviour at hand, we do not need to impute another capacity to explain it instead (e.g., Morgan's Canon). Whatever the merits of this view, this kind of parsimonious thinking may be less useful when trying to distinguish between rich and lean cognitive processes in humans – as we already have good reason to think that humans are in possession of *both* rich and lean cognitive capacities. Thus, what the cognitive scientist needs is not a strategy which allows them to characterise the broader ontology of cognitive mechanisms in humans, but a strategy that allows us to determine which kind of cognitive process is deployed in which setting. Our strategy – based on experimental design and contrastive testing – can serve this role.

with the advice frame should simply reflect how surprising a correct (or incorrect) recommendation is, averaging over all the recommendations in the recent past (e.g. has the frame been reliable or not?). In contrast, if the mentalist alternative is true, prediction errors for the red frame should become state-dependent. As participants, we will infer that if the last trial was successful, the advisor is likely to now be deceiving us (making a helpful recommendation improbable), while if the last trial was unrewarded they are likely to be cooperating (making an unhelpful recommendation more surprising).

The reason this alternative experiment provides an effective contrastive test is that it has the potential to yield patterns of results that are uniquely diagnostic of simple and complex learning algorithms. For instance, a diagnostic feature of prediction error learning algorithms of the kind deployed by Behrens et al. is 'win stay' behaviour. If a cue yields a reward, learners will be generally more likely to stick with that option than to switch – as experiencing reward increases the expected value of that option in future. But if agents mentalise, representing the advisor's strategy of making unhelpful recommendations after useful ones, we would instead expect learners to show 'win shift' behaviour – ignoring the social advice in a state-dependent way.

Crucially, the process of model development and fitting - as applied by Behrens and colleagues - cannot be used to provide this kind of contrastive test on its own. There is no way, in their experimental design, to obtain evidence that uniquely favours one of these accounts over another. This is because the rich and lean accounts do not predict any different empirical signatures. Under both possible theories, one source of information – one cue, such as the red frame - is used to make predictions, compute errors, and revise subsequent beliefs – which is all the algorithm mandates. To rule out the operation of more complex mentalising processes, it is necessary to create experimental conditions where these complex processes have the potential to yield different results.

The mentalist case

Recall that, in our view, mentalists have a different problem. They are on firm ground when they infer complex cognitive processes from the success of a complex over a simple model (Step 1 > Step 2), but there are difficulties when they infer social specificity from cognitive complexity (Step 2 > Step 3).

The mentalist's problem can be illustrated by a recent study by Velez and Gweon (2019, Experiment 3). In each trial in this experiment, two cards were randomly drawn from a set marked with values between 1 and 8. Initially, the value of one card was visible and the value of the other was hidden. Two groups of participants could choose to 'stay' and win the points on the visible card, or 'switch' and win the points on the hidden card. Before making this decision, they were advised to stay or switch by an agent, Alice, who had only seen the hidden card. Alice's advice was equally accurate for all participants – she recommended the correct action in more than 70% of trials - but Alice was 'Conservative' when advising one group, recommending 'stay' unless the hidden card had a very high value, and 'Risky' when advising the other group, recommending 'switch' unless the hidden card had a very low value. Across both groups, a more complex model explained more variance in stay-switch behaviour than a simpler model. The more complex model, called the 'Mental-state Reasoning model', calculated the value of the hidden card by combining the value of the visible card with Alice's advice, along with estimates that tracked whether Alice's recommendations were accurate or not, and whether her advice tended to be cautious or cavalier. The simpler model, an 'Accuracy Heuristic model', assumed a high value for the hidden card when the advisor recommended switching, a low value when she recommended staying, and weighted this advice by an estimate of the advisor's accuracy.

It is very tempting to conclude from this result, not only that the volunteers' stay-switch decisions were based on a cognitively complex process (Step 1 > Step 2), but that the complex process was socially specific, mental state reasoning (Step 2 > Step 3). That is certainly what the creators of the more complex model intended the 'Mental-state Reasoning model' to model, and it is entirely plausible on an introspective basis. We can imagine ourselves as one of the volunteers, staring at the visible card, and wondering about Alice's character or what she can see. Is she a bit reckless? And, of course, this tempting conclusion may well be correct. But the purpose of experiments like this is to go beyond introspection, to check whether our hunches are right – or, more precisely, the range of conditions in which they are right. We are not asking whether adult humans are capable of mental state reasoning. We know they are. Rather, we are asking whether adult humans sometimes, routinely, or always use mental state reasoning when learning from others.

The volunteers in Velez and Gweon's experiment completed more than a hundred decision trials in rapid succession, online, and with minimal financial incentives (a potential maximum of $2 for points earned in the whole experiment). Under these conditions, it is plausible that their cognitive systems used a short cut. They may have used the cues specified by the more complex model (the number on the visible card, the rectangle around 'stay' or 'switch' in this trial, how common it had been for the rectangle to surround 'stay' or 'switch') to estimate the value of the hidden card without representing this value as the content of another agent's mental state. Specifically, they may have calculated, not only the reliability of the rectangle cue (success probability), but also its bias; in the terms of signal detection theory, they may have calculated the propensity of the rectangle to 'miss' or register a 'false alarm'. This would explain the success of the complex model, but it does not guarantee that participants were mentalising the source of the rectangle. If people used the kind of shortcut we have in mind, the complex model would fit just as well if the 'Alice story' was replaced by an 'Alarm story'. The Alarm is programmed to tell you to switch when its camera detects a 'high value' card, but some alarms are triggered only by really high card values (Conservative), while others are sometimes triggered by relatively low card values (Risky). We could track the sensitivity and bias of such an alarm without attributing mental states.

A common response to this social specificity problem is to add 'neuro' to the cognitive science. Brain imaging can establish whether performance in a social learning task is correlated with, or causally influenced by, activity in areas of the brain associated with mental state reasoning – dorso-medial prefrontal cortex (dmPFC) and temporoparietal junction (TPJ) (Hampton, Bossaerts & O'Doherty 2008, Hill, Suzuki, Polania, Moisa, O'Doherty & Ruff 2017). If activity in these areas occurred always and only when people are engaged in mental state reasoning, or another socially specific cognitive process, this would be a good solution. But that is not the case. Current evidence leaves open the possibilities that processing in the dmPFC or TPJ is socially specialised, asocial (Lockwood, Wittmann, Apps, Klein-Flugge, Crockett, Humphreys & Rushworth 2018), or reflects a processing mechanism that is engaged, not only by mental state reasoning, but also by self-monitoring and metacognition (Frith 2012; Heyes, Bang, Shea, Frith & Fleming 2020; Lockwood et al. 2020).

A more reliable way of solving the social specificity problem is to use contrastive experiments with an 'inanimate' or 'non-social' control condition (Cook, den Ouden, Heyes & Cools 2014; Cook, Swart, Frobose, Diaconescu, Geurts, den Ouden & Cools 2018; Cook, Swart, Frobose, Diaconescu, Geurts, den Ouden & Cools 2019; Heyes 2014a, 2014b, Lockwood et al. 2020). Cook and colleagues have pioneered this approach in studies of social learning. Using an experimental task and modelling techniques like those of Behrens et al. (2008, see above), they told people that the red frame represented the most popular choice selected by a group of participants who had previously played the game (social group), or that it represented the outcome from a system of rigged virtual roulette wheels, which fluctuated between providing useful and less useful information (non-social group). Their research with this non-social control has revealed domain-specific processing but not for the social domain. They have shown that indirect learning differs from direct learning (Cook et al. 2018), and that meta-learning differs from first-order learning (Cook et al. 2019), but they have not found that different models fit the data when people believe they are getting information from other agents rather than rigged roulette wheels (Cook et al. 2018).

In contrast, Devaine, Hollard & Daunizeau (2014) found evidence of social specificity when they used a non-social control in a competitive, rather than a cooperative, advisory task. Confronting the same two-choice decision in every trial, participants were told that they were playing 'hide and seek' against a human opponent (social framing) or a 'casino gambling task' with two slot machines (non-social framing). In both cases, they were in fact playing against ‘artificial mentalising agents’ – programmes predicting the participant’s choice with varying degrees of sophistication. The participants were better at predicting their opponent’s choices in the social framing condition, and a complex model – intended to represent mentalising – fit the data better than simpler models when participants believed their opponent was another person but not when they believed they were playing against slot machines. These differences between the social and non-social conditions provide a strong indication, not only that the former recruited more complex cognitive processes, but that these processes were socially specific. The conclusion that mentalising was the socially specific process is less secure because, even in the social condition, participants with high scores on independent tests of empathy and executive function did not perform any better than people with lower scores. But this puzzle is not central to our current concerns. The important point is that Devaine et al. (2014) used a non-social control to overcome the problem that afflicts so many mentalist studies of social learning – securing the inference from cognitive complexity to social specificity (Step 2 > Step 3).

Can the results from Devaine et al. be used to shore up other work in the mentalist school? For example, can we infer that, because processing is socially specific in a hide and seek game it is also likely to be socially specific in the card game used by Velez and Gweon (2019)? This particular inference clearly would not be secure given that, using cooperative tasks similar to the card game, Cook et al. have tried and failed to find social specificity. More generally, to use the results from Devaine et al. as convergent evidence of social specificity, we would need to use non-social controls to find out much more about the range of conditions in which social learning depends on socially specific processes.

Scientific realism

We have suggested that two inferential problems haunt the use of computational models to find out about the complexity and social specificity of the cognitive processes mediating social learning, and that these problems can be solved by contrastive experiments. Our analysis assumes that computational cognitive scientists interested in social learning are scientific realists – that they want their work to yield true, or as-true-as-possible, descriptions of cognitive processes. Perhaps that assumption is wrong.

There are signs of an anti-realist historicism in research on social learning – of the view that associationist and mentalist theories are products of incommensurable paradigms (Kuhn 1970) and therefore cannot be tested against one another empirically. These signs often appear in developmental and comparative psychology. Mentalists who study social learning in children and nonhuman animals often imply that their view does not need to be tested against associative alternatives because associationism is a thing of the past; a view that went out of fashion with the behaviourist paradigm (Heyes 2012). In contrast, most computational cognitive scientists appear to be confident that empirical evidence can and should be used to compare associationist and mentalist theories of social learning. This is one of the great strengths of the field, one of the reasons why it has so much potential to advance our understanding of social learning.

Turning from historicism to instrumentalism, an anti-realist with logical positivist sympathies might argue that there is no problem inferring cognitive simplicity from model simplicity (Step 1 > Step 2) because computational models are merely instruments for predicting observable phenomena or categorising observations. If a model predicts what people (and their brains) will do when they have an opportunity to learn from others, it is a good model in the only way that a model can be good. There is no further question of whether the model maps onto or accurately describes unobservable cognitive processes. Moreover, if a simpler and a more complex model both predict social learning behaviour and brain activity, we are justified in preferring the simpler model on pragmatic grounds; because it is easier to work with. We need not argue that the simpler model is more likely to offer a true description of unobservable cognitive processes, or, therefore, to do contrastive experiments that give more complex processes an opportunity to shine.

Appeals to parsimony and the virtues of scientific unity could be signs of this kind of instrumentalism. When cognitive neuroscientists hail a model of social learning as parsimonious, they may see parsimony as a pragmatic virtue rather than an indicator that the model is likely to be true. When they point out that acceptance of a simple model would connect social learning with other kinds of learning in humans and other animals or that acceptance of a complex model would link social learning with other kinds of human causal reasoning (e.g., Velez & Gweon 2021), they may be assuming that unity of science is a virtue in itself, regardless of whether we have reason to believe that diverse phenomena are likely to be due to common underlying processes. As far as we are aware, these instrumentalist views, although apparent in research on social learning in artificial systems (Johanson, Hughes, Timbers & Leibo 2022; Leibo et al. 2022; Yaman, Leibo, Iacca & Lee 2022) have not been articulated in research on the cognitive neuroscience of social learning. Nonetheless, it is worth noting that the move from model simplicity/complexity to cognitive simplicity/complexity is problematic only if one is a true believer in cognitive processes. If talk about 'association', 'inference', and 'reasoning' is just a way of glossing a computational model, the true function of which is to predict brain activity and overt behaviour, the inference from Step 1 to Step 2 disappears. We are left only with questions about the clearest and most persuasive way of communicating model characteristics without equations.

If, as we suspect, the computational cognitive science of social learning is committed to scientific realism, why is there a tendency to neglect the kinds of experiments that are, on a realist view, needed to answer questions about the complexity and specificity of social learning? Some of the reasons are likely to be general. It is possible that, across all topics in computational cognitive neuroscience, experimental design tends to be weaker than in (other) areas of experimental psychology because practitioners need to find time, skills and funding for modelling, and often brain imaging, in addition to resources for behavioural experiments. Maybe, with so much to do, priority is given to the exciting new techniques involved in modelling and imaging, rather than the dull old ones that make a good behavioural experiment. In addition, historical currents may be having specific effects on the computational cognitive science of social learning. Although there is scant evidence of anti-realist historicism, associationists are part of a long tradition in which parsimony has been prized

(Thorndike 1898); a tradition that may incline them to overlook the need for experiments to distinguish simple and complex processes. Mentalists, on the other hand, are influenced by research on “theory of mind” which, since its inception (Premack & Woodruff 1978), has tended to assume that, if social behaviour is not due to a very simple kind of associative learning, it must be due to reasoning about mental states (Heyes 1998). With a blindspot around the possibility of domain-general reasoning about observable features of other agents, this tendency may obscure the need for experiments to find out whether complex processes are or are not socially-specific.

Conclusion

Computational cognitive neuroscience is opening the black box of social learning. After more than a century of research indicating that social learning has major evolutionary, developmental, and economic consequences, cognitive neuroscience is beginning to tell us how it works – what happens in an agent’s head when they are learning from others. Progress has been excellent but, we argue, two inferential problems have not yet been cracked. The problem of inferring cognitive complexity from model complexity calls for social learning to be tested in challenging task environments. The problem of inferring social specificity from cognitive complexity calls for non-social controls. More generally, cognitive neuroscientists need to combine their most distinctive and impressive tools – computational modelling and brain imaging – with subtle, contrastive experimental designs.

A yet more general moral of our story is that scientific realists in cognitive neuroscience, whether they study social learning or other functions, should work harder to keep track of where their models are supposed to be. Are they just scientific instruments, inferential tools in the heads and computers of the researchers? Or are they (also) natural phenomena, mental models in the heads of their participants? It is easy to slip from one to the other, to confuse the explanans with the explanandum, but realists need to separate models from minds.

Acknowledgments

We are grateful to Jennifer Cook for valuable discussion while we were writing this paper.

Competing Interests

The authors have no conflicting interests to declare.

References

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, 456(7219), 245-249.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214-1221.

Cook, J. L., Den Ouden, H. E., Heyes, C. M., & Cools, R. (2014). The social dominance paradox. *Current Biology*, *24*(23), 2812-2816.

Cook, J. L., Swart, J. C., Froböse, M. I., Diaconescu, A., Geurts, D. E., den Ouden, H. E., & Cools, R. (2018). Catecholamine challenge uncovers distinct mechanisms for direct versus indirect, but not social versus Non-Social, learning. *bioRxiv*, 303982

Cook, J. L., Swart, J. C., Froböse, M. I., Diaconescu, A. O., Geurts, D. E., Den Ouden, H. E., & Cools, R. (2019). Catecholaminergic modulation of meta-learning. *elife*, *8*, e51439.

Dacey, M. (2016). The Varieties of Parsimony in Psychology. Mind & Language, 31(4), 414-437.

Dacey, M. (2021). Evidence in Default: Rejecting default models of animal minds, British Journal of Philosophy of Science, p. 714799.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: does mentalizing make a difference when we learn?. *PLoS computational biology*, *10*(12), e1003992.

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology*, *30*(15), R860-R865.

Duhem, P. M. M. (1991). The aim and structure of physical theory, vol. 13. Princeton University Press.

Fitzpatrick, S. (2008). Doing away with Morgan's Canon. Mind & Language, 23, 224-246.

Fitzpatrick, S. (2017) Against Morgan's Canon. In The Routledge Handbook of Philosophy of Animal Minds, K. Andrews and J. Beck, Eds., Routledge.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213-2223.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896-910

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, *105*(18), 6741-6746.

Haselgrove, M. (2016). Overcoming associative learning. *Journal of Comparative Psychology*, *130*(3), 226.

Heesen, R., Bright, L. K. & Zucker, A. (2019). Vindicating methodological triangulation. *Synthese*, 196, 3067-3081.

Heyes, C. M. (1994). Social learning in animals: Categories and mechanisms. *Biological Reviews, 69,* 207-231.

Heyes, C. M. (1998). Theory of mind in nonhuman primates. Behavioral and Brain Sciences, 21(1), 101-114.

Heyes, C. (2012). Simple minds: a qualified defence of associative learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2695-2703.

Heyes, C. M. (2014a) False belief in infancy: a fresh look. *Developmental Science,* 17,647-659.

Heyes, C. M. (2014) Submentalizing: I'm not really reading your mind. *Perspectives on Psychological Science, 9, 131-143.*

Heyes, C. M. (2016). Blackboxing: social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society: B, 371,20150369.*

Heyes, C. M. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society:B, 373:20170051.*

Heyes, C. M., & Galef Jr, B. G. (Eds.). (1996). *Social learning in animals: the roots of culture*. Elsevier.

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences*, *24*(5), 349-362.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature neuroscience*, *20*(8), 1142-1149.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91-106.

Hoppitt, W., & Laland, K. N. (2013). Social learning: an introduction to mechanisms, methods and models. Princeton University Press.

Johanson, M. B., Hughes, E. Timbers, F. & Leibo, J. Z. (2022). Emergent bartering behaviour in multi-agent reinforcement learning. ArXiv220506760.

Leibo, J. Z., Köster, R., Vezhnevets, A. S., Duénez-Guzmán, E. A., Agapiou, J. P. & Sunehag, P. (2022). What is the simplest model that can account for high-fidelity imitation? Behavioral & Brain Sciences, vol. 45.

Yaman, A., Leibo, J. Z., Iacca, G. & Lee, S. W. (2022) The emergence of division of labor through decentralized social sanctioning'. ArXiv220805568.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. University of Chicago Press: Chicago.

Lockwood, P. L., Wittmann, M. K., Apps, M. A., Klein-Flügge, M. C., Crockett, M. J., Humphreys, G. W., & Rushworth, M. F. (2018). Neural mechanisms for learning self and other ownership. *Nature communications*, *9*(1), 1-11.

Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a ‘social’ brain? Implementations and algorithms. *Trends in Cognitive Sciences*, *24*(10), 802-813.

Macphail, E. M. (1985). Vertebrate intelligence: The null hypothesis. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 37-51.

Mobius, M., & Rosenblat, T. (2014). Social learning in economics. *Annu. Rev. Econ.*, 6(1), 827-847.

Olsson, A,. Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience, 21,* 197-212.

Pietraszewski, D., & Wertz, A. E. (2022). Why evolutionary psychology should abandon modularity. *Perspectives on Psychological Science*, *17*(2), 465-490.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10,* 59-63.

Poldrack, R.A., (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron, 72,* 692-697.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences*, *1*(4), 515-526.

Quine, W.V.O. (1958). Two dogmas of empiricism. In From a logical point of view, Cambridge, MA, US: Harvard University Press. pp. 20-46.

Reed, E.J., Uddenberg, S., Suthaharan, P., Mathys, C.D., Taylor, J.R., Groman, S.M., & Corlett, P.R., (2020). Paranoia as a deficit in non-social belief updating. *ELife, 9,* e56345.

Robbins, P. (2013). Modularity and mental architecture. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6), 641-649.

Roumazeilles, L., Schurz, M., Lojkiewiez, M., Verhagen, L., Schüffelgen, U., Marche, K., Mahmoodi, A., Emberton, A., Simpson, K. & Sallet, J. (2021). Social prediction modulates activity of macaque superior temporal cortex. *Science advances*, *7*(38), eabh2392.

Sober, E. (2009). Parsimony and models of animal minds. *The philosophy of animal minds*, 237-257.

Stanford, P. K. (2006). Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford University Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), i–109.

Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in Cognitive Science*, *11*(2), 299-315.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current opinion in behavioral sciences*, *38*, 110-115.

Zentall, T. R., & Galef, B. G. (Eds.). (1988). *Social learning: Psychological and biological perspectives*. Psychology Press.