PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Opinion piece

Cite this article: Heyes C. 2016 Blackboxing: social learning strategies and cultural evolution. *Phil. Trans. R. Soc. B* **371**: 20150369. http://dx.doi.org/10.1098/rstb.2015.0369

Accepted: 19 December 2015

One contribution of 15 to a theme issue 'Attending to and neglecting people'.

Subject Areas:
behaviour, cognition, evolution

Keywords:
asocial learning, associative learning, blackboxing, cultural evolution, metacognition, social learning strategies

Author for correspondence:
Cecilia Heyes
e-mail: cecilia.heyes@all-souls.ox.ac.uk

Blackboxing: social learning strategies and cultural evolution

Cecilia Heyes

All Souls College and Department of Experimental Psychology, University of Oxford, Oxford OX1 4AL, UK

Social learning strategies (SLSs) enable humans, non-human animals, and artificial agents to make adaptive decisions about *when* they should copy other agents, and *who* they should copy. Behavioural ecologists and economists have discovered an impressive range of SLSs, and explored their likely impact on behavioural efficiency and reproductive fitness while using the 'phenotypic gambit'; ignoring, or remaining deliberately agnostic about, the nature and origins of the cognitive processes that implement SLSs. Here I argue that this 'blackboxing' of SLSs is no longer a viable scientific strategy. It has contributed, through the 'social learning strategies tournament', to the premature conclusion that social learning is generally better than asocial learning, and to a deep puzzle about the relationship between SLSs and cultural evolution. The puzzle can be solved by recognizing that whereas most SLSs are 'planetary'—they depend on domain-general cognitive processes—some SLSs, found only in humans, are 'cook-like'—they depend on explicit, metacognitive rules, such as *copy digital natives*. These metacognitive SLSs contribute to cultural evolution by fostering the development of processes that enhance the exclusivity, specificity, and accuracy of social learning.

1. Introduction

Social interaction is not always a good thing. Even the members of a highly cooperative species, such as humans, are sometimes better off if they go it alone. Two heads can be better than one, but also too many cooks can spoil the broth. Some tasks call for the strength, skills, or knowledge of more than one agent, but the benefits of cooperation can be outweighed by the cognitive and emotional costs of coordination, or the incompetence of a cooperator. Sometimes other people just get in the way.

These truisms apply not only to 'joint action' [1], where agents' bodies move together to accomplish a common task, but also to 'informational cooperation' [2], where an agent typically acts alone but uses information derived from other agents, from 'social learning' or 'copying'.[1] In the last decade, cognitive scientists—including psychologists and neuroscientists—have focused more on joint action than on informational cooperation, and, as many of the articles in this Theme Issue show, made significant progress in identifying the mechanisms that allow second-by-second coordination of body movements. But they have not addressed systematically the broader question of when we should and should not engage in joint action; when it is more efficient to go it alone. By contrast, behavioural ecologists and economists have focused on informational cooperation rather than joint action, and discovered a great deal about the conditions in which it is and is not advisable to learn from others. Consequently, they have proposed that agents use 'social learning strategies' (SLSs), rules such as *copy when uncertain* and *copy if dissatisfied* [3] that specify the conditions in which it is prudent to engage in social rather than asocial learning. However, behavioural economists and ecologists have blackboxed the neurocognitive mechanisms involved in social learning, in asocial learning, and in switching between the two. Using what is known as the 'phenotypic gambit' [4], their research has probed the functional properties—effects on task efficiency and reproductive fitness—of SLSs, while remaining deliberately agnostic about the nature and origins of the neurocognitive mechanisms that implement these decision rules [5].

THE ROYAL SOCIETY PUBLISHING

Research on SLSs using the phenotypic gambit has been of considerable value in its own right, and provides an example for those who study joint action of how the costs and benefits of cooperation can be measured and modelled. We psychologists and neuroscientists would do well to take a similarly cool look at the pros and cons of joint action. However, in this article, I suggest that the approach adopted by behavioural ecologists and economists, blackboxing SLSs, is no longer a tenable scientific strategy. We need to open the black box, not merely for interdisciplinary piety, or to ensure that we understand the mechanisms as well as the functions of SLSs, but to avoid superficiality and error in the core domain—in characterizing the *functions* of SLSs.

To highlight the risks associated with the phenotypic gambit, I shall discuss two ideas that have gained currency in research on SLSs. The first idea is that when SLSs are applied, social learning is more efficient than asocial learning. In this case, I will argue that blackboxing has led to error. The second idea is that in humans, SLSs contribute to cultural evolution. In this case, I will argue that blackboxing has generated conflicting conclusions, and the conflict can be resolved only by thinking harder about both the cognitive mechanisms implementing SLSs, and the nature of cultural evolution.

But first, a caution: discussions of SLSs often make it sound as if social learning and asocial learning are mediated by different mechanisms—as if they are distinct 'gadgets', switched on and off according to the dictates of an SLS. This is misleading. In the past, when social and asocial learning were blackboxed in the way that SLSs are now, it was widely assumed that they are mediated by different mechanisms. However, there is now a substantial body of evidence that social learning differs from asocial learning only at the level of inputs; in the social case, the activity of another agent draws attention to, or instantiates, the cues that are learned [6]. The mechanisms that encode and store social and asocial inputs have been modelled as associative processes [7] and using Bayesian frameworks [8].

2. Is social learning better than asocial learning?

In 2010, Rendell *et al*. [9] published in *Science* the results of an open 'social learning strategies tournament' with a first prize of EUR 10 000. The real-world competitors were academic and non-academic computing experts. The virtual competitors were programmes they had written, each representing an SLS, a set of decision rules used by an individual agent. The programmes/SLSs/virtual agents competed with one another for points in a simulated 'multi-armed bandit' environment, where there were 100 different behavioural options, each with a different pay-off that varied probabilistically over time. In each round of the competition, the agent applied its decision rules—its SLS—to choose between three moves: Innovate, Observe, and Exploit. Rendell and colleagues defined these moves:

> Innovate represented asocial learning, that is, individual learning stemming solely through direct interaction with the environment, for example, through trial and error. An Innovate move always returned accurate information about the payoff of a randomly selected behavior previously unknown to the agent. Observe represented any form of social learning or copying through which an agent could acquire a behavior performed by another individual, whether by observation of or interaction with that individual. An Observe move returned noisy information about the behavior and payoff currently being demonstrated in the population by one or more other agents playing Exploit . . . Lastly, Exploit represented the performance of a behavior from the agent's repertoire, equivalent to pulling one of the multiarmed bandit's levers. Agents could only obtain a payoff by playing Exploit. ([9], p. 209)

The headline finding of the SLS tournament was that 'it is virtually always better to copy others than to figure things out for yourself' [10], i.e. social learning is better than asocial learning. This conclusion was based primarily on the data shown in figure 1. There we see the average final score of each of the 104 agents in the tournament, in pairwise competitions, plotted against the strength of each agent's bias towards Observe over Innovate moves. Thus, values on the *x*-axis, labelled 'Proportion of Observe when learning' were calculated by dividing, for each agent, the number of Observe moves by the number of Observe moves plus the number of Innovate moves. Figure 1 shows that success in the tournament was positively correlated with bias in favour of Observe over Innovate. However, if we open the black box, and think about the definitions of Innovate and Exploit in relation to psychology of learning, it becomes evident that this positive correlation does not provide evidence that social learning was more effective than asocial learning in the tournament.

Innovate was equated with asocial learning, but Innovate does not correspond with any form of asocial learning known to psychologists. Exploit was the only move said to involve 'the performance of a behaviour', and Exploit was contrasted with Innovate. This implies that Innovate did not involve performance, but how, in the real world, could an agent acquire a behaviour and error-free information about its pay-off without executing the behaviour, without acting on the world? Deduction, or an internal selection process [11], might inspire an agent with a new idea about what they could do, but it would not give them error-free information about the pay-off of this new behaviour. Similarly, a very smart friend might describe a new behaviour in detail and tell you the pay-off based on his or her performance of the behaviour, but this would be social, not asocial, learning.

Given that Rendell *et al*. defined Innovate as 'individual learning stemming solely through direct interaction with the environment, for example, through trial and error', it is unlikely that they meant to imply that Innovate did not involve action on the world. But if Innovate involved action execution as well as the receipt of information about pay-offs, what was the difference between Innovate and Exploit? From a psychological perspective, Exploit has all the hallmarks of 'trial and error' or 'reinforcement' learning, but in the tournament it was treated as an *alternative* to learning. This is very strange indeed because Exploit did not merely involve 'the performance of a behavior from the agent's repertoire, equivalent to pulling one of the multiarmed bandit's levers' ([9], p. 209). When an agent played Exploit, the executed behaviour sometimes yielded a pay-off that differed from the pay-off specified when the behaviour was added to the agent's repertoire (via Innovate or Observe), and in these cases the agent was able to 'update its knowledge of how profitable that act was, and store the updated information in its behavioral repertoire' ([9], Supporting Online Material). Therefore, whenever the agent used Exploit, it performed an action, experienced an outcome, and, when the outcome deviated from expectations, updated its record of the value of the action. Exploit was trial-and-error learning, red in tooth and claw.

Thus, the data in figure 1 do not provide evidence that social learning was more efficient than asocial learning in the SLS

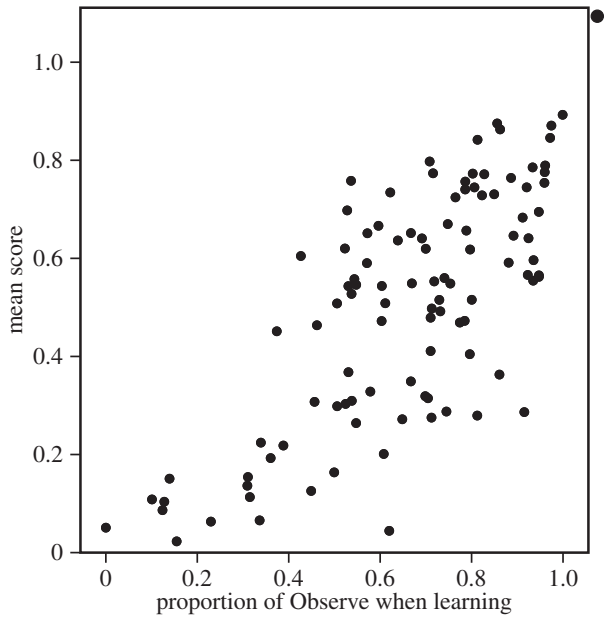

Figure 1. Mean final score in the social learning strategies tournament plotted against the proportion of Observe and Innovate moves that were Observe. Reproduced with permission from [9].

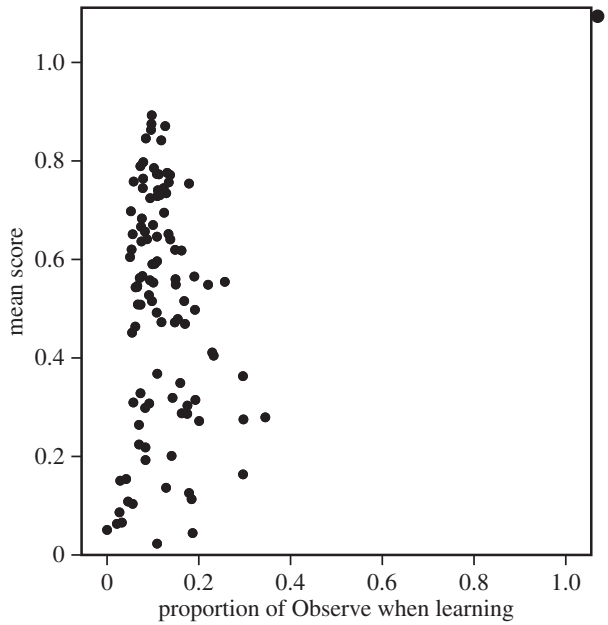

Figure 2. Mean final score in the social learning strategies tournament plotted against the proportion of all moves that were Observe, i.e. when Exploit was classified as learning.

tournament. They represent a black-box analysis that either equated asocial learning with processes for which there is no psychological evidence, or defined asocial learning such that it was virtually indistinguishable from Exploit—a move that any psychologist would recognize as trial and error learning, but which was mysteriously categorized in the tournament as not learning at all. In an attempt to get a more realistic picture of how social and asocial learning fared in the SLS tournament, figure 2 plots mean final score (the same *y*-axis as in figure 1) against the proportion of all moves that were Observe (Observe/total moves).[2] If we assume, with Rendell and colleagues, that Observe represents social learning and that Innovate is a form of asocial learning, and now classify Exploit also as a form asocial learning, then the *x*-axis in figure 2 indicates the extent to which each agent opted for social rather than asocial learning. The first thing to note about figure 2 is that all values are relatively low; the agents did not do much social learning at all. Furthermore, the positive correlation between Observe and final score, shown in figure 1, has disappeared. If anything, it seems that the few agents who did a relatively large amount of social learning were *less* successful than their competitors.

The picture painted by figure 2 is consistent with frequently overlooked empirical evidence that agents can be circumspect about the use of information from others, and that social learning can lead to incorrect decisions (see [12] for a review). However, figure 2 certainly does not show that asocial learning is better than social learning. To get reliable information about the relative value of social and asocial learning, it would be necessary, at minimum, to perform a deeper reanalysis of the SLS tournament data; for example, to divide Exploit moves into those that tested behavioural options derived from Innovate and from Observe. But it would be better still to design a new tournament based on psychologically realistic assumptions.

I have discussed the SLS tournament in some detail because it shows that opening the black box is necessary, not merely for completeness—to ensure that we understand the mechanisms as well as the function of SLSs—but to avoid mistaken conclusions about function [13]. To yield reliable information about function, mathematical approaches do not need to model psychological processes in detail. However, they do need to differentiate psychological processes in ways that are consistent with empirical research in cognitive science, to cut the mind at its joints.

3. How do social learning strategies contribute to cultural evolution?

(a) The problem

In addition to using computational modelling to explore the properties of SLSs (see previous section), behavioural economists and ecologists have conducted experiments to find out which SLSs are used by real agents, and how effective they are in producing adaptive behaviour. A significant proportion of this work has involved non-human animals. For example, Webster & Laland [14] allowed minnows, small freshwater fish, to feed at two locations, A and B, and then encouraged them to observe a shoal of conspecifics feeding at just one of these locations, B. Finally, they recorded which location the observers approached when tested individually under low, medium or, high predation risk. The minnows tested under high risk—when a model of a predator was present—were more likely than minnows tested under low or medium risk to copy the shoal, i.e. to show a preference for the location, B, at which they had seen conspecifics feeding. This result was taken to indicate that minnows use the SLS *copy when asocial learning is costly*.

Research of this kind has provided evidence that animals from a wide range of species—including rats, bats, fruit flies, and sticklebacks, as well as minnows—use a variety of SLSs, specifying *when* to copy (e.g. *copy when uncertain*, *copy when personal information outdated*), and *who* to copy (e.g. *copy older individuals*, *copy the majority*; reviews [5,15,16]). This work is important because it helps to explain why social learning is so widespread in the animal kingdom. If animals engaged in social learning indiscriminately, copying both more and less knowledgeable conspecifics equally often, it is unlikely that,

on average, social learning would be advantageous. However, the wealth of evidence that animals use SLSs also creates a problem. It is difficult to reconcile with the commonly held view that SLSs support human culture [3,17,18]. If many animals use SLSs, and SLSs promote culture, why is culture uniquely human? Why do not rats, bats, fruit flies, and sticklebacks show cultural diversity—geographical variation in behaviour that is not due to genetic or ecological differences? Why do not these species, equipped with SLSs, show cumulative cultural change—the accretion of wisdom over generations, through social learning, to produce sophisticated technology, elaborate social practices, and vast libraries of knowledge about the world [19]?

(b) The memory hypothesis

This problem has been largely ignored by behavioural ecologists and economists. Even the fullest and most persuasive discussion I have been able to find suggesting that memory plays a crucial role in enabling human SLSs to support culture is brief and indirect [18]. The structure of the argument is difficult to trace, and it is not presented as an overarching hypothesis about the relationship between SLSs and culture. Nonetheless, since Fogarty and colleagues are the only behavioural ecologists or economists to have offered something akin to a solution, I shall try to summarize the parts of their argument for which they cite evidence, and call this summary 'the memory hypothesis': unlike the SLSs of other animals, human SLSs make extensive use of memory. They keep track of a long history of pay-offs, and use this record, discounting older information, both to detect change in pay-offs and to estimate the probability that such a change is about to occur. Owing to these features, human social learning can be deployed with unparalleled efficiency; for example, only when the agent's own information is likely to be outdated and/or another agent is very likely to know better. This increase in efficiency 'allowed humans to invest in more sophisticated social learning than is seen elsewhere in the animal kingdom' ([18], p. 241).

On the surface, this is a plausible way of explaining why it is only human SLSs that promote culture, but it has three shortcomings. (i) *Evidence*. The only evidence that memory is important in distinguishing human from non-human SLSs comes from the SLSs tournament [18]. This is a problem, not only because the agents in the tournament were not really choosing between social and asocial learning (see previous section), but also because the memory-related analysis of the tournament data was post hoc and informal. After the tournament had been run, Fogarty *et al*. [18] divided the agents into five 'loose memory categories', and reported that agents assigned to the highest category—the ones judged to have made the most use of memory—had fared better in the competition than agents assigned to the other four categories. However, they did not explain on what basis they had decided that the programmes in the highest category used memory 'to estimate environmental parameters', and 'to predict the probability of certain environmental changes in the future or discounting'. Furthermore, and more important, they did not cite any empirical evidence that real live humans do, and real live animals do not, use memory in these ways to support SLSs. (ii) *Blackboxing*. References to 'memory' and 'sophisticated social learning' do not really open the black box; they just lift the lid a centimetre or two, and then let it slam shut. What are the memory processes that enable estimation of environmental parameters, prediction of future changes, and discounting? What is 'sophisticated' about sophisticated social learning, and how, at the level of neurocognitive mechanisms, does it differ from unsophisticated social learning? To explain what is distinctive about human SLSs, a memory hypothesis would need to answer these questions, or direct us very firmly to a body of research that can provide the answers. (iii) *Causality*. As it stands, the memory hypothesis does not offer an account of how extensive use of memory could enable human SLSs to promote cultural evolution. It mentions a potential mediator, sophisticated social learning, but does not explain how extensive use of memory would allow 'humans to invest in more sophisticated social learning than is seen elsewhere in the animal kingdom', or how the additional sophistication afforded by extensive use of memory would enable human SLSs to support cultural evolution.

(c) The metacognition hypothesis

A recently proposed alternative to the memory hypothesis suggests that the crucial difference between humans and other animals is that *some* of the SLSs used by humans are explicitly metacognitive. Only these SLSs have the characteristics that have previously been ascribed, by implication, to all SLSs: they are reportable, domain-specific rules that represent 'who knows', i.e. properties of the cognitive processes of the rule user and of other agents [19]. Potential examples of metacognitive SLSs include *copy the boat builder with the largest fleet* and *copy digital natives* (see below). Metacognitive SLSs focus social learning on knowledgeable agents so precisely that they encourage high-fidelity copying of behaviour. Because it is exclusive, specific and accurate, this kind of copying promotes cultural evolution by enhancing 'parent–offspring relations' [20], keeping useful changes small, and encouraging their proliferation to many agents.

Unlike the memory hypothesis, the metacognition hypothesis is rooted in cognitive science. Consequently, although far from complete or fully confirmed, the metacognition hypothesis has more empirical support than the memory hypothesis; opens the black box much wider; and offers a more detailed account of how (some) human SLSs contribute to cultural evolution.

(i) Evidence

Careful examination of comparative data suggests that animal SLSs are based exclusively on relatively simple, domain-general cognitive processes, such as associative learning [15]. These processes are domain-general in that they operate in the same way when processing inputs from social and asocial sources. They are certainly products of genetic evolution, but they did not evolve specifically to guide social learning.

As an example, let us consider the minnows mentioned at the beginning of this section. These minnows showed a stronger tendency under high predation risk, than under low or medium risk, to approach a feeder, B, at which they had seen a shoal of conspecifics feeding [14]. In principle, this modulation of copying by risk could be due to a psychological process that is dedicated to 'gating' socially acquired information; that opens the gate, allowing such information to control behaviour, when predation risk or other indicators of costly asocial learning are high, and closes the gate when they are low. However, the modulation effect can also be explained by domain-general processes of associative learning, and this alternative explanation is supported by other

studies of SLSs in animals [15,21]. According to this account, initial training had two effects. First, through appetitive conditioning, a form of asocial learning, the minnows developed a tendency to search for food at locations A and B. Second, through social learning, the minnows acquired a tendency to approach location B, where conspecifics had previously been seen to shoal. There is abundant evidence that the rate at which an animal responds for food can be weakened considerably if an aversive stimulus, such as a cue for shock, is presented. This conditioned suppression effect has been observed in a variety of species, including goldfish [22]. As the level of threat increases, therefore, it is likely that it will exert a similar, suppressive effect and reduce any tendency to search for food in A, or anywhere else, and thus allow the tendency to approach B, solely because of its association with a shoal of conspecifics, to manifest itself fully [15].[3]

Many human SLSs are also likely to be mediated by domain-general cognitive processes [23]. However, there is evidence that some of the SLSs used by adult humans are explicitly metacognitive; that people are deciding whether or not to learn from others on the basis of conscious, reportable rules specifying *when* other agents are likely to have superior knowledge, and *who*, among the available models, is likely to know best [19]. For example, in foraging and perceptual tasks, people were asked to make a preliminary decision, and an explicit judgement of their confidence in that decision, before being given the opportunity to use social information to make a final decision [24]. The participants' confidence judgements were accurate—they had lower confidence in wrong than right preliminary decisions—and, crucially, they were increasingly likely to use social information as their confidence declined, suggesting that they deliberately applied the rule *copy when uncertain*. Similarly, in another foraging task, people made use of social information—advice about which of two options to choose—to the extent that they believed the advisor to be motivated to help rather than to mislead them [25]. These beliefs were explicitly stated, and the basic effect—covariation between the advisors' incentives and the participants' use of social information—disappeared when participants were told that the advisors did not know which option they were recommending. Therefore, these results indicate that the participants used an explicitly metacognitive strategy such as *copy when the model intends to help*.

Altogether, the currently available data on SLSs are consistent with a dual-systems account in which most SLSs are 'planetary', based on domain-general psychological processes, while a few, found only in humans, are 'cook-like', based on explicit, domain-specific rules [19]. Planetary motion conforms to rules, but planets do not understand these rules or implement them deliberately; the rules of planetary motion are in the minds of scientists, not in the minds of planets. Similarly, the behaviour of animals can be described and predicted by SLSs, but the strategies or rules are in the minds of scientific observers, not of the animals themselves. By contrast, when people use explicitly metacognitive SLSs, they are like cooks rather than like planets. Cooks know the rules to which their behaviour conforms, and the conformity of their behaviour is due, in part, to their knowledge of the rules.

(ii) Opening the black box

Explicit metacognition is one of the most complex phenomena tackled by cognitive science. Within dual-systems models of the mind [26,27], and related theories [28,29], explicitly metacognitive representations are part of a cognitive system that handles problems slowly and serially. Its functioning depends on working memory, and correlates with differences between people in general intelligence. By contrast, implicit metacognition, and many of the cognitive processes represented by explicit metacognition—including the domain-general processes underpinning SLSs in animals—are part of a cognitive system that handles problems rapidly and in parallel, and that is minimally dependent on working memory.

Explicit metacognition is typically studied by cognitive scientists using judgements of learning and confidence judgements. When working towards an examination, students use explicitly metacognitive judgements of their own prior learning to exclude from future study materials they have already assimilated, and to prioritize material they have nearly, but not quite, mastered [30,31]. When making perceptual decisions—for example, about the presence or orientation of an object in an array—people use explicitly metacognitive confidence judgements to decide how much they should bet on the accuracy of their decisions, and to communicate the reliability of their decisions to cooperation partners [32–34].

Rapid progress has recently been made in understanding the neural and computational bases of metacognition [34], and in research on its development. The latter suggests that explicitly metacognitive rules are learned [33,35]; this learning typically depends on social interaction [32,36,37]; and consequently, there is marked cultural variation in explicit metacognition [38–41]. For example, children learn by instruction to use 'semantic clustering' to retrieve the names of animals from memory ([36]; e.g. think of birds first and then mammals), and adults learn through social interaction explicitly to metarepresent their confidence in ways that make two heads better than one [32,37].

Given these findings from research on metacognition in general, one would expect that if some human SLSs are explicitly metacognitive, they should also be products of learning through social interaction, and vary across cultures. Evidence consistent with the first of these predictions indicates that the SLSs found in pre-school children, like those found in animals, depend on domain-general processes [23]. This suggests that domain-specific SLSs do not emerge until relatively late in development, when there has been ample opportunity for them to be learned through social interaction. There is also evidence supporting the second prediction, that there will be marked cross-cultural variation among the SLSs used by adults [42–46]. For example, in contrast with Westerners, Fijians are *less* likely to seek advice from people with more formal education [45], and, in contrast with Britons, people from mainland China engage in more social learning, and their social learning is less dependent on uncertainty [42].

(iii) Causality

The dual-systems account of SLSs is very far from deflationary. It suggests that even planetary SLSs are much more flexible and efficient than was previously thought. If, as behavioural ecologists and economists have assumed, SLSs were fixed products of genetic evolution, SLSs would make social learning selective, but only in a way that was efficient in ancestral environments. For example, if older individuals had tended to provide more reliable information in the distant past, agents alive today would be inclined to *copy*

older individuals even if, in a tech-savvy world, younger individuals tend to know more, at least on certain topics. By contrast, because they are rooted in domain-general processes of learning, planetary SLSs can make social learning selective in a way that is adjusted rapidly, within lifetimes, to track changes in the social and asocial environment. Thus, if younger individuals provide more reliable information in particular contexts, agents will learn to attend to and copy younger more than older individuals in those contexts.

Because it casts even planetary SLSs as extraordinarily supple and adaptive, the dual-systems view makes it especially hard to identify the advantages of cook-like SLSs, and thereby to spell out what it is about (some) human SLSs that enables them to promote cultural evolution. The metacognition hypothesis tries to meet this challenge in the following way.

Step 1: Metacognition to better SLSs. Explicitly metacognitive SLSs are able to focus social learning on knowledgeable agents with greater accuracy and precision because metacognitive SLSs are themselves products of cultural evolution. Acquired through learning in the context of social interaction (see 'Opening the black box'), metacognitive SLSs distil the accumulated wisdom of many agents about when and which others know best. Planetary SLSs can be updated on the basis of only one agent's experience—the user's experience. If an agent gets higher pay-offs when she copies younger individuals than when she sticks to asocial learning or copies older individuals, she will develop through domain-general mechanisms—by planetary means—a bias to *copy younger individuals*. But this bias only has a modest chance of being adaptive because it is narrow, derived from a relatively small sample of younger and older individuals—the small number of individuals that the focal agent has tried copying. By contrast, when a middle-aged person learns, by explicit instruction or via the zeitgeist, to *copy digital natives*, she is acquiring a metacognitive SLS based on the pay-off experience of a large number of other people—including all those who have been educated about information technology by their children, and have let this be known to others. Thus, whereas genetically inherited SLSs would be broad but inflexible, and planetary SLSs are flexible but narrow, metacognitive SLSs are both broad and flexible.

Step 2: Better SLSs to higher fidelity. Because metacognitive SLSs identify 'who knows' with greater accuracy and precision, they increase the likelihood that agents will gain more by copying with higher than lower fidelity. In this context, fidelity has at least three components: (i) Exclusivity—deriving information from one or a small number of models, rather than by combining information from a large number of other agents (e.g. *copy the majority*). (ii) Specificity—copying at a fine rather than a coarse grain- exactly when, where, how, and in what order small components of the action are performed. (iii) Accuracy—copying without introducing random error, or changes based on asocial learning [47]. When high-fidelity copying is at a premium, metacognitive SLSs promoting exclusivity are favoured by cultural evolution (e.g. *copy the boat builder with the largest fleet*, will gain more currency than *copy the majority's boat* design [20]), and both individual agents and social groups can afford to invest in the development of tools and cognitive mechanisms that allow copying with high specificity and accuracy. The cognitive mechanisms include executive processes focusing attention on the details of a model's behaviour, and encoding the serial order of its components and sensorimotor processes enabling translation of what has been observed into matching action by the observer [48,49].

Step 3: Higher fidelity to cultural evolution. Godfrey-Smith has shown that models of cultural evolution seek to explain the *distribution* of cultural traits—for example, why some ideas or skills are more common than others in certain social groups—and/or the *origin* of cultural traits—for example, how particular skills, such as building a canoe from seal skin, could possibly come into existence [20]. The success of distribution explanations—of applying models from population genetics to cultural phenomena—depends on there being good parent–offspring relations in the cultural domain. As in gene-based evolution, each new instance or 'token' of a cultural type must be a copy of one or a small number of existing tokens. It is not sufficient for the earlier-occurring and later-occurring tokens merely to be alike, or for the latter to be loosely inspired by the former [20,50,51]. My bread-making skill is the offspring of your bread-making skill to the extent that I acquired my skill by copying your technique, and resisted blending your technique with others I observed, or with my own bright ideas about bread making. Therefore, if metacognitive SLSs promote exclusivity and accuracy in social learning (see Step 2)—if they reduce the number of models contributing to each new token of a cultural trait, and the degree to which the model's influence is contaminated by asocial learning—metacognitive SLSs will enhance parent–offspring relations, and thereby increase the power of population genetic models to explain the distribution of cultural traits. Or, to make the same point more directly: metacognitive SLSs help to create the conditions in which the distribution of cultural variants can evolve geographically over time.

Origin explanations—of how improvements in a cultural variant could accumulate to produce something as impressive as a seal skin canoe—are not as dependent on strong parent–offspring relations as distribution explanations. However, they do require that cultural variants change in small steps, and that useful new variants proliferate through the population in a way that creates many 'independent platforms for further tinkering' [20]. In other words, an impressive achievement can be ascribed to cultural evolution, rather than to the insight and ingenuity of a succession of individual agents, to the extent that each improvement was made more likely by there being *many* agents, rather than *smart* agents, using its precursor [52]. Therefore, both the specificity and the accuracy of social learning are relevant to origin explanations. Specificity—copying at a fine rather than a coarse grain—helps to keep innovations small, and accuracy—copying with a minimum of random error, or changes based on asocial learning—helps to ensure that small innovations proliferate intact to many agents within the population. There is always tension in Darwinian evolutionary models between variant generation and faithful retention [11], but to the extent that metacognitive SLSs support detailed and accurate copying, they are likely to help cultural evolution, rather than the smart choices of a succession of individual agents, to produce complex theories, artefacts, and practices.

In summary: I have argued that because it does not blackbox cognitive mechanisms, the metacognition hypothesis is better able than the memory hypothesis to explain not only the psychology of SLSs, but the function of SLSs—how planetary SLSs contribute to the development of adaptive behaviour in general, and how cook-like, metacognitive SLSs contribute to cultural evolution. In combination with signs that blackboxing led the SLSs tournament mistakenly to conclude that social learning is better than asocial learning, this discussion suggests that the phenotypic gambit should no longer be used in research

on SLSs. To find out why and how agents target their social learning, we need to combine the resources of behavioural ecology, behavioural economics, *and* cognitive science. We need to open the black box good and wide.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

Acknowledgements. I am grateful to Vincent Crawford, Peter Godfrey-Smith, John Pearce, Luke Rendell, Nick Shea, and an anonymous referee for their comments on an earlier draft of this article.

Endnotes

[1]By convention, 'copying' is used as a synonym for 'social learning', in spite of the fact that social learning by agent A from agent B can result in A's behaviour being *less* similar to B's behaviour than it would have been if A had not learned from B. For example, if I learn that an object is hot by observing you touch it and wince, I will be less likely to touch the object than if I had not observed your behaviour. Thus, the conflation of social learning with copying may itself be a consequence of blackboxing—of failure to consider the psychological mechanisms that mediate social learning, and therefore to recognize that they can yield systematically non-matching as well matching behaviour. By contrast, research on joint action recognizes that agents can use similar or dissimilar behaviours to accomplish a task together.

[2]I am very grateful to Luke Rendell, not only for providing the data shown in figure 1*b*, but also for offering to provide any data from the SLS tournament, and to assist with further analysis.

[3]The associative account predicts that Webster & Laland [10] would have obtained the same result if in the second phase of the experiment they had allowed the minnows to observe a small cave (an asocial stimulus), rather than a shoal of conspecifics, at location B.

References

1. Sebanz N, Bekkering H, Knoblich G. 2006 Joint action: bodies and minds moving together. *Trends Cogn. Sci.* **10**, 70–76. (doi:10.1016/j.tics.2005.12.009)
2. Sterelny K. 2012 Language, gesture, skill: the co-evolutionary foundations of language. *Phil. Trans. R. Soc. B* **367**, 2141–2151. (doi:10.1098/rstb.2012.0116)
3. Rendell L, Fogarty L, Hoppitt WJE, Morgan TJH, Webster MM, Laland KN. 2011 Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends Cogn. Sci.* **15**, 68–76. (doi:10.1016/j.tics.2010.12.002)
4. Grafen A. 1984 Natural selection, kin selection and group selection. In *Behavioural ecology* (eds JR Krebs, NB Davies), pp. 62–84. Oxford, UK: Blackwell Scientific Publications.
5. Hoppitt W, Laland KN. 2013 *Social learning: an introduction to mechanisms, methods, and models*. Princeton, NJ: Princeton University Press.
6. Heyes C. 2012 What's social about social learning? *J. Comp. Psychol.* **126**, 193–202. (doi:10.1037/a0025180)
7. Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. 2008 Associative learning of social value. *Nature* **456**, 245–249. (doi:10.1038/nature07538)
8. Perreault C, Moya C, Boyd R. 2012 A Bayesian approach to the evolution of social learning. *Evol. Hum. Behav.* **33**, 449–459. (doi:10.1016/j.evolhumbehav.2011.12.007)
9. Rendell L *et al.* 2010 Why copy others? Insights from the social learning strategies tournament. *Science* **328**, 208–213. (doi:10.1126/science.1184719)
10. Rendell L. 2011 Culture evolves: the social learning strategies tournament. Bute Media Lab. See https://www.youtube.com/watch?v=q22eaqmzQW8.
11. Campbell DT. 1974 Evolutionary epistemology. In *The philosophy of Karl Popper* (ed. PA Schilpp), pp. 413–463. La Salle, IL: Open Court.
12. Rieucau G, Giraldeau L-A. 2011 Exploring the costs and benefits of social information use: an appraisal of current experimental evidence. *Phil. Trans. R. Soc. B* **366**, 949–957. (doi:10.1098/rstb.2010.0325)
13. Lewens T. 2015 *Cultural evolution*. Oxford, UK: Oxford University Press.
14. Webster MM, Laland KN. 2008 Social learning strategies and predation risk: minnows copy only when using private information would be costly. *Proc. R. Soc. B* **275**, 2869–2876. (doi:10.1098/rspb.2008.0817)
15. Heyes C, Pearce JM. 2015 Not-so-social learning strategies. *Proc. R. Soc. B* **282**, 20141709. (doi:10.1098/rspb.2014.1709)
16. Laland KN. 2004 Social learning strategies. *Anim. Learn. Behav.* **32**, 4–14. (doi:10.3758/BF03196002)
17. Laland KN, Rendell L. 2013 Cultural memory. *Curr. Biol.* **23**, R736–R740. (doi:10.1016/j.cub.2013.07.071)
18. Fogarty L, Rendell L, Laland KN. 2012 Mental time travel, memory and the social learning strategies tournament. *Learn. Motiv.* **43**, 241–246. (doi:10.1016/j.lmot.2012.05.009)
19. Heyes C. In press. Who knows? Metacognitive social learning strategies. *Trends in Cognitive Sciences*. (doi:10.1016/j.tics.2015.12.007)
20. Godfrey-Smith P. 2012 Darwinism and cultural change. *Phil. Trans. R. Soc. B* **367**, 2160–2170. (doi:10.1098/rstb.2012.0118)
21. Galef Jr BG, Dudley KE, Whiskin EE. 2008 Social learning of food preferences in 'dissatisfied' and 'uncertain' Norway rats. *Anim. Behav.* **75**, 631–637. (doi:10.1016/j.anbehav.2007.06.024)
22. Geller I. 1964 Conditioned suppression in goldfish as a function of shock-reinforcement schedule. *J. Exp. Anal. Behav.* **7**, 345–349. (doi:10.1901/jeab.1964.7-345)
23. Heyes C. In press. When does social learning become cultural learning? *Dev. Sci.* (doi:10.1111/desc.12350)
24. Morgan TJH, Rendell LE, Ehn M, Hoppitt W, Laland KN. 2011 The evolutionary basis of human social learning. *Proc. R. Soc. B* **279**, 653–662. (doi:10.1098/rspb.2011.1172)
25. Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, Fehr E, Stephen KE. 2014 Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput. Biol.* **10**, e1003810. (doi:10.1371/journal.pcbi.1003810)
26. Csibra G, Gergely G. 2006 Social learning and social cognition: the case for pedagogy. In *Processes of change in brain and cognitive development Attention and performance XXI*, vol. 21 (eds Y Munakata, MH Johnson), pp. 249–274. Oxford, UK: Oxford University Press.
27. Kahneman D. 2011 *Thinking, fast and slow*. New York, NY: Macmillan.
28. Norman DA, Shallice T. 1986 Attention to action: willed and automatic control of behaviour. In *Consciousness and Self-Regulation* (eds RJ Davidson, GE Schwartz, D Shapiro), pp. 1–18. New York, NY: Springer.
29. Dehaene S, Naccache L. 2001 Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37. (doi:10.1016/S0010-0277(00)00123-2)
30. Metcalfe J. 2009 Metacognitive judgments and control of study. *Curr. Dir. Psychol. Sci.* **18**, 159–163. (doi:10.1111/j.1467-8721.2009.01628.x)
31. Yue CL, Castel AD, Bjork RA. 2013 When disfluency is—and is not—a desirable difficulty: the influence of typeface clarity on metacognitive judgments and memory. *Mem. Cogn.* **41**, 229–241. (doi:10.3758/s13421-012-0255-8)
32. Mahmoodi A, Bang D, Ahmadabadi MN, Bahrami B. 2013 Learning to make collective decisions: the impact of confidence escalation. *PLoS ONE* **8**, e81195. (doi:10.1371/journal.pone.0081195)
33. Timmermans B, Schilbach L, Pasquali A, Cleeremans A. 2012 Higher order thoughts in action: consciousness as an unconscious re-description process. *Phil. Trans. R. Soc. B* **367**, 1412–1423. (doi:10.1098/rstb.2011.0421)
34. Fleming SM, Dolan RJ, Frith CD. 2012 Metacognition: computation, biology and function. *Phil. Trans. R. Soc. B* **367**, 1280–1286. (doi:10.1098/rstb.2012.0021)
35. Shea N, Boldt A, Bang D, Yeung N, Heyes C, Frith CD. 2014 Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18**, 186–193. (doi:10.1016/j.tics.2014.01.006)

36. Hurks PPM. 2012 Does instruction in semantic clustering and switching enhance verbal fluency in children? *Clin. Neuropsychol.* **26**, 1019–1037. (doi:10.1080/13854046.2012.708361)
37. Bahrami B, Olsen K, Bang D, Roepstorff A, Rees G, Frith C. 2012 Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 3. (doi:10.1037/a0025708)
38. Heine SJ, Kitayama S, Lehman DR, Takata T, Ide E, Leung C, Matsumoto H. 2001 Divergent consequences of success and failure in Japan and North America: an investigation of self-improving motivations and malleable selves. *J. Pers. Soc. Psychol.* **81**, 599–615. (doi:10.1037/0022-3514.81.4.599)
39. Mayer A, Träuble BE. 2013 Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *Int. J. Behav. Dev.* **37**, 21–28. (doi:10.1177/0165025412454030)
40. Li J. 2003 US and Chinese cultural beliefs about learning. *J. Educ. Psychol.* **95**, 258. (doi:10.1037/0022-0663.95.2.258)
41. Güss CD, Wiley B. 2007 Metacognition of problem-solving strategies in Brazil, India, and the United States. *J. Cogn. Cult.* **7**, 1–25. (doi:10.1163/156853707X171793)
42. Mesoudi A, Chang L, Murray K, Lu HJ. 2015 Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of cultural evolution. *Proc. R. Soc. B* **282**, 20142209. (doi:10.1098/rspb.2014.2209)
43. Toelch U, Bruce MJ, Newson L, Richerson PJ, Reader SM. 2014 Individual consistency and flexibility in human social information use. *Proc. R. Soc. B* **281**, 20132864. (doi:10.1098/rspb.2013.2864)
44. Eriksson K. 2012 The nonsense math effect. *Judgment Decis. Mak.* **7**, 746–749.
45. Henrich J, Broesch J. 2011 On the nature of cultural transmission networks: evidence from Fijian villages for adaptive learning biases. *Phil. Trans. R. Soc. B* **366**, 1139–1148. (doi:10.1098/rstb.2010.0323)
46. Efferson C, Richerson PJ, McElreath R, Lubell M, Edsten E, Waring TM, Paciotti B, Baum W. 2007 Learning, productivity, and noise: an experimental study of cultural transmission on the Bolivian Altiplano. *Evol. Hum. Behav.* **28**, 11–17. (doi:10.1016/j.evolhumbehav.2006.05.005)
47. Goodnow JJ. 1955 Determinants of choice-distribution in two-choice situations. *Am. J. Psychol.* **68**, 106–116. (doi:10.2307/1418393)
48. Catmur C, Walsh V, Heyes C. 2009 Associative sequence learning: the role of experience in the development of imitation and the mirror system. *Phil. Trans. R. Soc. B* **364**, 2369–2380. (doi:10.1098/rstb.2009.0048)
49. Watson JS. 1972 Smiling, cooing, and 'the game'. *Merrill-Palmer Q. Behav. Dev.* **18**, 323–339.
50. Sperber D. 2000 An objection to the memetic approach to culture. In *Darwinizing Culture: the status of memetics as a science*, pp. 163–174. Cambridge, UK: Cambridge University Press.
51. Shea N. 2009 Imitation as an inheritance system. *Phil. Trans. R. Soc. B* **364**, 2429–2443. (doi:10.1098/rstb.2009.0061)
52. Amundson R. 1989 The trials and tribulations of selectionist explanations. In *Issues in evolutionary epistemology* (eds K Hahlweg, CA Hooker), pp. 413–432. New York, NY: State University of New York Press.