The Triumph of a Reasonable Man: Stich, Mindreading and Nativism


Version of January, 2004

Kim Sterelny
Philosophy Program, Australian National University and Victoria University of Wellington

kimbo@coombs.anu.edu.au; Kim.Sterelny@vuw.ac.nz
1. Beyond Simulation and The Theory-Theory.

Humans interpret others. We are able to anticipate both the actions and intentional states of other agents. We do not do so perfectly, but since we are complex and flexible creatures even limited success needs explanation. For some years now Steve Stich (frequently in collaboration with Shaun Nichols) has been both participant in, and observer of, debates about the foundation of these capacities (Stich and Nichols 1992; Stich and Nichols 1995). As a commentator on this debate, Stich (with Nichols) gave explicit and fair-minded sketches of the cognitive architectures presupposed by the various theories of mindreading. As a participant, Stich has mostly been a defender of the theory-theory, the view that normal human agents have an internally represented theory of other agents and they use that theory in interpreting other agents. The main recent rival to this position, simulationism, claims that agents use their own decision-making mechanisms as a model of those of other agents, and derive their predictions by modelling others in something like the way aeronautical engineers derive predictions from the use of scale models in wind-tunnels. Stich has been sceptical about this alternative, for on his view simulation theory makes mistaken predictions about both the development of interpretive competence and about the pattern of interpretive success and failure.

In their most recent work, Stich and Nichols have shifted ground, developing an impressive simulationist/theory-theory hybrid (Nichols and Stich 2004). I do not intend to evaluate this hybrid model in this paper. Rather, I intend to explore its consequences for nativist conceptions of interpretation. Under the influence of the linguistics model, nativist ideas have become increasingly prominent in cognitive and evolutionary psychology. Applied to mindreading, nativism emerges as the idea that folk psychological concepts, folk psychological principles, or both, are innate. They develop early, universally and independently of any informational contribution from experience. Folk psychology is developmentally entrenched. In recent work, I have been increasingly sceptical of this nativist conception of our interpretive capacities (Sterelny 2003b; Sterelny forthcoming), and I shall argue that if Stich and Nichols are right about the architecture of mindreading, these sceptical arguments are reinforced.
This conclusion would not be very exciting if their view was hopelessly implausible. But in my view, a hybrid approach to the problem of understanding our interpretive capacities has a good chance of being true. Simulationist ideas are made plausible by the suspicion that the theory-theory over-intellectualises interpretation; this is particularly evident from the fact that relatively young children (five year olds) are quite competent interpreters. If we interpret by learning and applying a theory of mind, how come we can do this so young? Perhaps our early faculty with a folk psychological theory can be explained by supposing that it is innate. But prediction does not come through theory alone: data and auxiliary hypotheses are needed. To apply innate principles to a particular case, they must be combined with a rich and accurate data set. For even if our theory of others is itself innate, if the theory-theory is true, interpretation requires a considerable data base, so for it early competence remains a puzzle. Yet simulation cannot be the whole story, for we are sometimes able to allow not just for differences in beliefs and preferences in predicting others, but also for different cognitive capacities. Good poker players regularly exploit weaker ones fallible probabilistic reasoning. Prima facie, interpretation seems likely to be the result of some mix of (as they put it) “information-poor” and “information-rich” processes. So the consequences for nativism of a hybrid model of interpretation are important. First, however, an outline of that model.

2. A Hybrid Theory of Interpretation and Prediction

The Stich-Nichols conception of mind-reading begins with a picture of an intentional system without metarepresentational capacities. The mind of this hypothetical agent includes a belief system with Updater and Planner: these are inference mechanisms taking inputs from and to the belief box. The UpDater, as the name suggests, updates the belief box in response to new information, whereas the Planner takes goals as inputs and gives as output a proposed course of action to realise that goal. The belief and desire box feed into a practical-reasoning decision-making system which in turn feeds into action control. What would we need to add to this agent to make it an interpreting intentional system: an agent with beliefs about the acts and thoughts of other agents? On the Stich-Nichols view, first, our interpreting agent needs to be sensitised to the goals of other agents. With this, a minimal form of interpretation is already possible.
Having attributed a desire to another agent, our interpreting agent can use her own beliefs and her own Planner to determine a strategy through which the interpretive target can achieve his goal, and she can then predict that the target will exploit that strategy. Of course, these interpretive capacities are quite limited even if the interpreting agent is a sophisticated detector of others’ goals, for there is no recognition of the differences in belief between interpreter and target. So such an interpreter will often mispredict the target's actions, for a strategy she would use will sometimes depend on a conception of their world the target does not share.

Stich and Nichols suggest that more sophisticated interpretive capacities depend on co-opting a cognitive capacity that is likely to be present for other purposes: a capacity for representing and reasoning about hypothetical situations. Suppose, for example, we are worried about the consequences for our academic program if the department chair were to resign. To reason about such scenarios we have (they suggest) a dedicated workspace, a “Possible World Box”. When we reason hypothetically, and all goes well, the hypothetical is “switched on” in the Possible World Box. Thus we write “The chairman has resigned” into the Possible World Box and lock that proposition on. We then add those of our beliefs that are consistent with that supposition to this workspace and then UpDate in the Possible World Box. The result is our prediction of what would happen should the chair resign.

In the Stich-Nichols view, the Possible World Box is co-opted to account for the differences in the target’s beliefs and the interpreter’s beliefs. The interpreter loads the Possible World Box with the her best estimate of the target's beliefs, and she uses the contents of the Possible World Box and the Planner to determine the target’s likely actions. In their view, importantly, a specific heuristic guides this procedure of estimating the target’s beliefs. The default procedure is to suppose that the interpreter’s beliefs and the target’s beliefs are the same. This default explains why we can routinely attribute common knowledge and its obvious implications, and why interpretation is often quite successful with strangers. But though this is the default procedure, it is overridden in particular cases. We use information about differences in perceptual perspective (“she could not see that from there”), information from conversations and other social interactions, and information from the agent’s behaviour to fine-tune our
estimate of the target’s beliefs. So, for example, we infer that she could not have realised that the fish is poisonous, for otherwise she would not have picked it up. There are many common desires, too, (especially avoidance-desires), so a parallel argument suggests that our default is to suppose that others have the same desires that we do. Just as we know without direct evidence that our target will believe that a crow is smaller than a whale, equally we know without direct evidence that Jane will not want to go to work with her ears painted blue. Despite this apparent parallel, this default-desire heuristic does not play a central role in the Stich-Nichols picture of desire estimation. In their view, in its earlier developmental states, we read desire off other agents’ expression, orientation and action. The fact that Peter is reaching for the beer in plain sight in front of him makes it easy to infer that he wants a beer.

We explain our own actions intentionally, too, and Stich and Nichols have a suggestion about first person interpretation, one which does not depend on an information-rich process. Instead, our minds are equipped with a special purpose detector that registers a belief in the belief box (or a preference in the preference box) and which writes a further, metabelief in the belief box. This loop involves no theory of mind or nor other information-rich understanding of intentional states. Detecting one’s own current beliefs is just a self-monitoring process, one which has the effect of adding a belief to the belief box. It is a brute-causal process, and is relevant only to self-awareness: it plays no direct role in understanding others. Monitoring your own thoughts is not at all like recognising the thoughts of others. Reasoning about your own thoughts, on the other hand, is like reasoning about those of others. Reasoning about thought is symmetrical. Working out what you would want if (say) your current desires were satisfied uses the same mechanisms as those involved in the counterfactual predictions of others’ desires.

This theory of mind reading is a hybrid. The utility of the default attribution of the interpreter’s belief set to the Possible World Box depends on interpreter-target similarities. So too does the use of the interpreter’s own Planner and UpDater to update the target’s doxastic world and to estimate the target’s course of action once her desire world has been identified. However, as Stich and Nichols point out, there are many elements of their view of mind-reading which do not depend on predictor-agent
similarities. Reading desire from expression and orientation does not: someone who loathes beer can zero in on Peter’s thirst for a beer as easily as a fellow beer-lover. Likewise, the mechanisms which filter and override default attribution do not depend on interpreter-target similarities.

3. Innateness on a Hybrid Model

Interpretive capacities develop very robustly: like language, the capacity to interpret others is a feature of all normal human minds. The pattern of development is similar from child to child, though there is variation in the rate at which the pattern unfolds (Peterson and Siegal 1999). Moreover, it is arguable that the information to which children are exposed is not rich; that is, the information on which interpretation rests is not manifest in the primary social experience to which children have access. Thus on some views the unobservability of psychological states makes it impossible for children to learn about the mental causes of behaviour. On this view, so-called “poverty of the stimulus” considerations show that folk psychology cannot be learned. A further and quite striking fact is that there is significant insensitivity to quite marked differences in general learning capacity: normal children and children with moderate retardation pass false belief tests at similar ages. In short: developmental patterns do not vary from individual to individual, even when there are marked differences in learning abilities; moreover, development proceeds in an informationally impoverished environment, and so the crucial principles connecting intentional states to action must be innate. In the light of these facts it is no surprise that interpretation has been assimilated to the language model, and regarded as the result of an innate module.

However, despite the prima facie plausibility of modular nativism, I have argued that these facts have an alternative explanation. The robust, stable development of interpretive capacities is the result of environmental scaffolding based on perceptual cues. For in addition to our perceptual responsiveness to facial expression and the like, children live in environments in which mutual interpretation is ubiquitous and made public by language. Moreover, the development of our capacities to understand other agents is supported by other social tools: for example, novels, plays and stories
provide people with a set of intentional templates (Sterelny 2003b; Sterelny forthcoming).

There is an important convergence between the anti-nativist arguments I have developed and the Stich-Nichols conception of mind-reading. On both our views, the development of mind reading is stabilised by quasi-perceptual mechanisms (by “shallow” modules). These play a crucial role in desire-reading: that is, in detecting emotional expression, the focus of attention, orientation, and the differences between deliberate and accidental motion. Shallow modules play an important role in estimating the doxastic world of other agents by factoring-in the role of differences in perceptual points of view. It is also stabilised by the co-option of cognitive machinery that evolved for other purposes. The Updater, the Planner and the Possible World Box are not specialisations for mindreading, but their co-option for mindreading stabilises this developmental trajectory. For central modes of reasoning do not have to be built from scratch to understand other agents. We are indeed perceptually adapted to mind-reading: we have perceptual mechanisms tuned to the focus of another’s desires and to differences in perceptual points of view. But mind-reading reasoning depends more on co-opted systems that systems built from scratch to interpret others.

In view of these convergences, my aim is to parasitise the Stich-Nichols view: to show that if their conception of our mind reading capacities is broadly correct (for nothing in the arguments that follow depend on the fine-grained details of their position) the anti-nativist case is thereby strengthened. I shall discuss three issues. I shall begin with poverty of the stimulus considerations. I shall then discuss the insensitivity of interpretive capacities to differences in general learning ability. Finally, I shall argue that innate adaptations and learned automatised skills are subject to different constraints, and that interpretation has the marks of a system constrained by limits on learning, not limits on adaptation.

---

1 A system whose importance I have previously overlooked. It is probably an ancient one, since the great apes are sensitive to these facts: (Call 2001).
4 The Poverty of the Stimulus

The poverty of the stimulus argument for an innate folk psychology has been sketched rather than defended in detail. So, for example, Leslie and Scholl have argued:

“As such, a ToM has often been thought to require its owner to have acquired the concept of belief. This raises the obvious developmental question: how do we acquire such a concept, and thereby acquire ToM. The obvious empiricist reply — that we learn it, perhaps from our parents — is suspect, due to the extremely abstract nature of mental states. After all, it’s not as if a parent can simply point and say ‘See that’ That’s a belief’. Children become competent reasoners about mental states, even though they cannot see, hear or feel them.” (Scholl and Leslie 1999) p133

Despite its brevity, this line of thought is widely seen as plausible. In considering it, let me begin with the familiar distinction between concept innateness and information (or knowledge) innateness. For this argument is most plausibly taken to suggest that intentional concepts cannot be learned from experience. I shall argue to the contrary that that there is no special difficulty in supposing that intentional concepts are learned. On my view, the crucial nativist issue is whether information needed for mindreading is innate. Nothing in this argument depends on the Stich-Nichols picture, but I shall then co-opt that picture to argue that if their view of the informational resources needed for mind-reading is right, those resources can be accumulated through ordinary learning mechanisms. The Stich-Nichols view of mindreading undercuts a nativism about mindreading based on poverty of the stimulus arguments.

So let’s begin with the idea that intentional concepts are innate. If we concentrate on belief, there is a picture of concept acquisition that lends plausibility to this

---

2 For example, in the discussion of this argument at the “Culture and the Innate Mind” conference in Sheffield, 2003, Botteril, Segal and Carruthers all defended it.

3 For the sake of argument, I shall proceed as if there is a clear distinction between concept nativism and informational nativism, but in fact I doubt that such a distinction can be drawn. First, it is important to note that those who accept nativist ideas about concept possession nonetheless accept that experience is necessary for concept acquisition. Thus innate concepts have to be “switched on” by experience. Thus the distinction between a learned concept and an innate one is typically made by appealing to a distinction between types of experiential cause; for example a distinction between an informationally sensitive learning process and mere triggering by experience (Fodor 1981; Cowie 1998). Yet the experience needed to trigger a concept is experience of instances of that concept. If the concept TIGER is triggered by experience it is triggered by tigers, not buttercups. Given that concepts, if triggered, are triggered by their instances, the learning-triggering distinction becomes extremely problematic if we further suppose (as Fodor and others do) that the possession of a concept is atomistic.
suggestion, a picture that sees concept acquisition as some kind of stimulus generalisation from instances. Thus experiences of particular pigeons leads to the development of a pigeon prototype which in turn is the basis of (or is perhaps even identical to) a concept of pigeons. It is indeed very hard to see how a concept of belief could be acquired by any route like this. However, while the whole issue of concepts and their possession is deeply opaque, we should be very cautious in buying an innateness hypothesis on the basis of any line of thought like this. For a large class of human concepts cannot be learned by stimulus generalisation: consider such representative examples as concepts for such kinds as: vehicle, tool, lawyer, father, fungus, food. Moreover Fodor is right, I think, to doubt that such concepts as these are definable: so they cannot be learned by specifying necessary and sufficient conditions of their application. Unless we are prepared to embrace the ultra-nativism of (Fodor 1975), we should not infer from “not learnt from instances by stimulus generalisation” to “innate”.

To be fair, one might reject a stimulus-generalisation view of concept learning and still wonder whether intentional concepts could be learned from experience. For if we focus on belief, it is true that there is seems to be a large gap between the activity that

That is, it does not involve propositional content of any kind (Fodor 1998). On this atomist view, having the TIGER-concept does not involve having a theory of tigers. Likewise to have the BELIEF-concept is to have a cognitive mechanism that responds to beliefs; it is not constituted by knowing stuff about belief. If any version of an atomistic view of concept possession turns out to be right, the distinction between learning and triggering looks set to collapse. For concept acquisition, if triggered, is triggered by instances. Yet unless concept possession has an informational component, in what way would such a stimulus be information-poor?

Thus given that concepts are triggered by their instances, conceptual atomism does not seem to support a distinction between triggered and learned concepts. We can draw that distinction if we drop atomism, but in doing so we erode the distinction between concept nativism and information nativism. For example, on descriptivist or theoretical-role views of concept possession, the possession of a concept is tied to an appropriate informational resource. That enables us to draw a triggering/learning distinction. If, for example, possession of the TIGER-concept is constituted by having a certain set of true beliefs about tigers, we can indeed ask whether the triggering experiences of tigers were informationally rich enough for the agent to have acquired from them that information. Along similar lines, if the possession of intentional concepts is tied to information about intentional states and their role in cognition, we can indeed ask whether there was enough accessible information in the signal for the agent to be able to learn enough about (say) belief from that signal. But then concept nativism fades into nativism about folk psychological principles. For example, if an agent can only have the concept of belief if she realises that beliefs are truth apt, that they are evidentially tied to perception and that they combine with preferences in practical inference, then concept nativism becomes a version of nativism about folk psychological principles. For the agent would have to represent those principles in order to acquire those concepts.

4 Thus Scott Atran appealed to this model of concept acquisition in developing a parallel argument for the innateness of the species concept: see (Atran 1998)
an agent can observe and the concept of belief. An agent’s beliefs are not distinctively manifest in his or her behaviour. Beliefs contribute causally to many actions, but they do not often have distinctive, readily learnable behavioural signature. My belief that there is beer in the fridge will take me to the fridge only if I want a beer and only if that desire is not trumped by more overriding concerns. This point is familiar from functionalist critiques of behaviourism, but though in general the behavioural expression of an agent's psychological state is modulated by her other psychological states, notice that belief stands at the extreme end of this interaction spectrum. Consider how belief contrasts, say, with disgust or embarrassment. While these states have no strict behavioural definition, they have reliable behavioural manifestations and so, to a greater or lesser extent, do many other psychological states. For this reason, folk psychology contains an array of iceberg concepts: concepts which name a syndrome that includes both a mental state and its distinctive behavioural manifestation. Desire (in its hot, affective sense); our basic emotions; some bodily sensations and perhaps some perceptual states (for example the focus of visual attention) are semi-observable psychological states, hence are named by iceberg concepts of this kind.

With iceberg concepts, the intuitive gap between observable activity and our concepts for the mental causes of that activity is less wide. You can point to the distinctive manifestation of an itch. Moreover, an agent who has mastered the concept of an itch has mastered the concept of an internal cause of action. The same is true of more obviously cognitive mental states. For example, children take time to master how to disguise their desires and to voluntarily delay their gratification. It is typically very easy to read their current desire off their actions, expression and perceptual focus. The same is true of men in bars. In an important sense, we can often point at desires. Moreover, if such iceberg concepts can be learned, their acquisition can facilitate the acquisition of concepts for less overt states. For they prime an agent to the possibility of internal causes of action. Children do seem to acquire the ability to reason about desires earlier than they acquire the ability to reason about belief. In the light of this argument, that fact is very suggestive. Iceberg concepts are a ramp for the acquisition of concepts for more cryptic mental states.
We lack a good general theory of the nature and acquisition of concepts, but even so, I doubt that there is a special, intractable problem of learning intentional concepts, despite the unobservability of intentional states. So I shall focus on the idea that we have innate information about intentional states and their roles. And here the Stich-Nichols model helps greatly. On the picture they paint, the information that agents need to interpret others is information that could well be acquired through learning. To begin with, the Stich-Nichols model shares the benefit of any model incorporating significant simulationist elements. In comparison to theory-theory views, there is less you need to know to read minds. If Stich and Nichols are right, to understand other agents, we do not have to represent the process by which other agents plan to achieve their goals at all. Moreover, and most importantly, the Stich-Nichols model reinforces a non-nativist conception of the development of mind reading by decomposing the construction of doxastic and desire worlds into subcomponents each of which can be both acquired and improved more or less independently of the others. Even on their hybrid picture, interpreting others is an information-rich task. But the information needed is of a kind that can be acquired and upgraded piecemeal.

Consider, for example, default belief attribution. In constructing the target’s doxastic word, you load your own beliefs into your Possible World Box, and then modify that belief set, cutting some elements and adding others. In the earlier stages of developing interpretive skills, this revision may be guided by nothing more than sensitivity to different perceptual histories. But at some point agents learn to use linguistic information: people talk about their own beliefs and argue and gossip about the beliefs of others. There are other social cues: our target will be treated with deference on some issues but not others. Thus if our target is treated as an expert on rugby, we will credit her with a rich and accurate set of football beliefs. Action, too, is informative: once we have learned to run practical syllogism in reverse, an agent’s actions tells us a good deal about her beliefs. If we see a distinguished and rather staid professor on his hands and knees with his head under the table, we can infer that he believes he has lost something there. In other circumstances, we might infer that an agent has found something there. Much the same is true of the identification of an agent’s goals. The crucial point is that these capacities can all be built one by one: they are not a package deal. Moreover, each component can be improved gradually.
For example, we gradually become better at reading subtle social cues to another agent's beliefs and preferences. We fine-tune reverse practical inference by learning to constrain the space of likely motives and beliefs which would lead to the actions we actually see. We learn to be better at disregarding pseudo-information, at ignoring exaggeration and malicious gossip.

Poverty of the stimulus arguments are persuasive when the representations of a domain develop accurately despite very noisy data: hence linguistic versions of the poverty of the stimulus argument have sometimes emphasised the ubiquity of performance errors in children’s linguistic experience. They are persuasive when development proceeds despite severe and relevant limitations in the data: hence linguistic versions of the poverty of the stimulus argument have often emphasised the lack of negative data in children’s linguistic experience. They are persuasive when development proceeds despite the fact that the representations are very unobvious given the data, for the concepts needed to describe the data are very different from those needed to predict and explain it. Hence linguistic versions of the poverty of the stimulus argument have often emphasised the abstractness of syntactic principles and their highly indirect relation to the data that confirm them (Cowie 1998). In explaining mind-reading, Stich and Nichols posit a mix of “information rich” and “information poor” mechanisms. But the information needed by the information rich mechanisms does not have a noisy, or a limited or a cryptic relationship to the experience from which the information could be learned.

5 Acquisition and General Learning Abilities.

Nativism about mindreading seems to be strikingly confirmed by the result that moderately handicapped children pass false belief tests near enough on schedule, for this suggests that mindreading is independent of general learning capacities (Scholl and Leslie 1999; Leslie 2000a; Leslie 2000b). I shall suggest an alternative possibility: the apparent independence of mindreading capacities from general learning capacities is an artefact of the concentration on the early development of mindreading, when shallow quasi-perceptual modules are indeed playing the central role in filtering default belief attribution and in desire attribution. If this is correct, we
would expect to see over time, as children age, an increasing disparity between the mindreading skills of normal and handicapped agents.

So here is my alternative scenario. Early in their mind-reading career, normal children depend on shallow modules to detect the desires of other agents. And even after they can exploit their “Possible World Box” to allow for the difference between their beliefs and those of the interpretive target, their default attribution strategy is simple. They attribute their own beliefs to the target, except those beliefs ruled out by differences in perceptual situation. Thus they can allow for the fact that they (but not the target) saw what was inside the Smarties box but in other respects they expect the agent to share their doxastic world. One subsequent developmental change is the acquisition of an increasingly sophisticated and increasingly learning-dependent set of skills for assessing the desires of another agent. Agents begin by using the perceptual cues of facial expression and intentional movement (and perhaps default attribution of one’s own desires). But over time they learn to use language, other social information, and inference from action, and these become increasingly important. As I discussed above, each of these individual components of skill can itself be incrementally improved. A parallel developmental change is an increasingly sophisticated and increasingly learning-dependent set of skills for overriding default belief attribution. Here too language, other social information, and inference from action also become increasingly important.

So this scenario predicts that as mind-reading becomes (i) increasingly dependent on upgraded mechanisms of desire identification and (ii) increasingly dependent on greater sensitivity to the differences between the interpreter’s beliefs and the target’s beliefs, the mind-reading gap between normal and learning-handicapped agents will become ever greater. The apparent insensitivity of interpretive skills to differences in learning capacity is an artefact of choosing a life stage at which these skills are relatively underdeveloped: it is an artefact of taking the false belief test to be too much of a watershed. If mindreading is really language-like in having its cognitive basis in a module, the innate-module view presumably makes the opposite prediction: it predicts that the disassociation between mindreading capacity and general learning
capacity is a more permanent feature of the mental life of learning-handicapped agents.

6. Optimality, Adaptation, Learning

The argument so far has co-opted the Stich-Nichols picture of the cognitive architecture of mindreading to argue against nativist views of that ability. I turn now to other elements of their picture: their emphasis on the limits as well as the strengths of our mindreading skills. And (it must be said) the argument becomes much more speculative. We are good but not perfect interpreters of others. Successful coordination typically depends on our being able to anticipate the actions and thoughts of our partners, and we often do succeed in co-ordinating, even when we must match actions that will take place months in the future. But there are also significant and systematic gaps in our capacity to anticipate others. Some of these gaps are no mystery. For example, we are much less sure of our interpretation of those agents from other cultures. For the mechanisms that filter our default belief and preference attribution depend on contingent, local information that we do not have. Our imperfect interpretation in such cases is no surprise on the Stich-Nichols view, but it is no surprise on anyone else’s view either. Everyone realises that particular belief-preference profiles are sensitive to individual history.

Other gaps in our capacity are potentially more informative, and Stich and Nichols discuss two such blindspots. First, perception. In some respects we are very sensitive to information about perception. In determining another agent’s goals, and in calibrating the differences between their doxastic world and ours, we take into account the focus of perceptual attention and differing points of view. But in other respects we are blind to facts about perception. The discovery of perceptual illusions is surprising precisely because of the existence of such blindspots. Thus Stich and Nichols discuss the wonderful example of change-blindness: if you show subjects two scenes within which there is an unchanged natural focus of attention in the
foreground, many will not notice quite significant background alterations. We do not anticipate perceptual illusions hence cannot anticipate their doxastic effects.

Stich and Nichols discuss a second type of case: context effects on decision making. The famous Milgram experiments were so shocking because they revealed powerful, unanticipated and socially dangerous context effects. Many experimental subjects in a social context that included an authority figure were prepared, despite their worries and hesitancies, to inflict what they believed to be dangerous electric shocks at that authority figure’s behest. This sensitivity of judgement and action to context has subsequently been demonstrated in many other contexts. One of the most entertaining of these used subjects from a theological college and probed the willingness of the subject to stop and help somehow who had (apparently) collapsed from illness. A striking result from these experiments was the sensitivity of willingness to help to time pressure: subjects running late to give a lecture (on, of all things, the fable of the good Samaritan) were much less willing to help than those not so pressured (Darley and Batson 1973).

These discoveries are striking because they are so unanticipated: they show not just irrationalities in human reasoning, they show irrationalities to which mindreading is blind. We are good at working out what others will think and do, but we are not optimal. This gap between optimal and actual interpretive capacities is potentially informative. For the hypothesis that interpretation is a learned automatised skill (albeit one that piggybacks on perceptual adaptations) and the hypothesis that interpretation is an innately-structured adaptation make different predictions about the gap between actual and optimal interpretation. If our interpretive capacities are based on a set of innate adaptations, we would expect to see two kinds of systematic failure. First, we would not expect to see accurate prediction in biologically irrelevant contexts, unless prediction in those contexts is a by-product of prediction in relevant ones. Second, information-rich but innate responses to the environment are vulnerable to environmental change. Suppose, for example, that determining what other agents

---

5 There is a wonderful video that illustrates a similar effect. You show a subject a scratch basketball game and assign quite a demanding task: count the number of times the ball is passed. In that context, many subjects (and I was one) will not notice such a dramatic event as a man in a gorilla suit walking through the game.
want is an informationally demanding task, and we succeed by having an innate data base about the desires people have and about how those desires manifest themselves in action and expression. We recognise that longing look, because we have its template wired into us for comparison. A system of this kind is vulnerable to environmental change. It will become more error prone if there are shifts in what people want or in how their desires are manifest in action. Wiring-in crucial information is good design if the environment is constant, for learning can be costly and vulnerable to the accidents of developmental environment. But it is good design only if the environment is constant.

So innately based systems will be far from optimal in a changing world. On the other hand, innate systems escape an epistemic constraint on individual learning: selection can see a pattern in noisy data because it sums over the life experience of many individuals. Evolution works over the timespan of a whole lineage, not the experience of an individual organism. Suppose that when agents lie, they have a slight but real tendency to stutter or slur at the beginning of their fabrication. Even an agent with inferentially optimal techniques for learning about her world may have no way of recognising this signal from her own experience. There will be honest stutters, and lies with perfect diction. She herself may experience more of the former than the later. Even if that is not the case, if the covariation is not marked, she will have no way of determining whether her own experiences are evidence of a systematic pattern. But if there is in the population a difference in levels of trust in response to such forms of language, selection can detect this small but significant difference. Indeed, the defenders of evolutionary psychology have seized on this idea as part of their general case for the adaptive superiority of special purpose innately structured systems over general purpose learning:

"many relationships necessary to the successful regulation of action cannot be observed by any individual during his or her lifetime. .... Domain general architectures are limited to knowing what can be validly derived by general purpose processes from perceptual information". ((Cosmides and Tooby 1995) p. 92)

Of course, it is true that innate systems can generate adaptive response to variation in the environment by wiring-in conditional response. However, there are limits to such wired-in conditional responses. The range of variation and its significance must itself be constant.
In short, if we assess other agents’ doxastic and desire worlds using largely innate though information-rich mechanisms, we would expect environmental change to degrade the accuracy of these assessments but we would expect to be alert to stable but noisy signals. If mind reading depends on learned, automatised skills, we make the opposite prediction. Environmental change should not degrade our mindreading capacities, but we might well fail to discover the value of real but noisy signals, especially as our general learning mechanisms are far from perfect in using noisy data.

Which of these predictions is confirmed? The crucial case here seems to be the effects of social context. Prima facie, Milgram effects, bystander effects and the like are of profound biological significance. For decision-making in social contexts is ubiquitous. Why then are our preference-detection mechanisms relatively insensitive to this information? This is especially puzzling because it is the strength of these effects, not their existence, to which we are blind. Thus, to take one of the more entertaining studies Stich and Nichols discuss (pp 136-137), it was no surprise to the experimental subjects that some students were prepared to defend in public an examination system they detested to please a pretty girl. It was the prevalence of this effect that the subjects failed to predict. Prima facie, this is a feature of the human mind onto which a innate, special purpose adaptation should lock. We need only suppose that interpreting agents have varied somewhat in their cynicism about the robust attachment of others to principle. Given such variation, those who expected—more than their contemporaries — conformist effects would have reaped the rewards of their superior predictive talents, and the population would have been gradually tuned to the right degree of cynicism.

Of course, not too much can be made of this suggestion, even if the premises are right. Adaptations can fail to be optimal for any number of reasons. But if our knowledge of the operating principles of other minds is innate, our blindness to the importance of social context is somewhat surprising. If, however, we have to learn how others think and act, it is not so surprising. Even if our general learning capacities were faultless, the strength of context effects would be quite hard to learn
from ordinary social interaction. To find out about this feature of the human mind, an interpreter would have to have good information about an agent’s preferences\textsuperscript{7} prior to the social context in question; good information about how those preferences change or are overridden in that social context; and the interpreter would need to be able to calibrate this effect quantitatively. That is, an interpreter would need this information about enough agents and situations to detect both the general trend and its strength. The task is objectively tough, and we are far from perfect reasoners about statistical information. It took good experimental design to reveal these effects about our decision making propensities. So it is no surprise that we did not discover them reasoning imperfectly about natural experiments.

So my hypothesis is that insensitivity to the effects of social context is the result of a noisy learning environment. We would have to learn about these context from imperfect data, and so our relative blindness to these effects can be naturally explained by appeal to constraints on learning. But equally (as Peter Godfrey-Smith has pointed out to me) they may be the effects of change. Human social worlds have been transformed in the last 10,000 years or so, and perhaps context effects are a consequence of these changes. So an alternative possibility is that our innate interpretive capacities have not caught up with the changed world in which they now operate. I think there are empirical tests (albeit difficult ones) which could discriminate between these two explanations of context blindness. We should test for context effects in environments more typical of those of our ancestors: that is, in social worlds which are smaller in scale, more egalitarian, culturally homogenous, and with little movement in and out. These social worlds are far from anonymous: most agents know one another well. If context effects are due to rapid environmental change, these effects should be less marked in such environments. The shaman’s mask should be less effective than the doctor’s labcoat in inducing Milgram-style conformity to authority. Likewise, we should test for the effects of informational enrichment. We should provide subjects with plenty of opportunities to learn about context effects and see whether this information becomes incorporated into automated, routine social judgement. If our blindness to context effects is due to rapid environmental change rather than a noisy learning environment, informational

\textsuperscript{7} Or beliefs, for there are context effects on judgement, too.
scaffolding will not help much. On the other hand, if scaffolded learning explains the robust and early development of our typical interpretive skills, scaffolding the effects of context should help greatly. I have my hunch about how such tests would turn out, but a hunch is not evidence.

7. Reprise

Let me briefly recapitulate the argument of this paper. Our ability to mindread is a cognitively sophisticated capacity that nonetheless develops robustly and early in all normal agents. In those respects it is like language, and thus it is no surprise that it too has been taken to depend on rich, innate information about its domain. I have argued that we can co-opt the Stich-Nichols picture of mindreading to undermine this nativist suggestion. For it undermines "poverty of the stimulus" arguments for mindreading by showing that the informational resources needed for mindreading decompose into separable elements, each of which can be acquired and improved independently of the others. This acquisition is promoted by genuine perceptual adaptations for mindreading: we are perceptually tuned to such salient features of other agents as their direction of gaze, their expression, and their orientation. These features of mindreading (if the Stich-Nichols picture is right) help explain a second puzzling fact: mindreading skills seem to be independent of general learning abilities. I have suggested that this appearance may be deceptive. It is generated by the critical role of perceptual adaptations — which are indeed independent of general learning abilities — in the first stages of the development of mindreading. If this suggestion is right, learning plays a central role in upgrading the default strategies for belief and desire attribution, and so in older children and adolescents mindreading skills and learning ability should vary together. Finally, I have argued that the limits on our mindreading capacities are more plausibly explained by the limits on individual learning than by the limits on adaptive evolution.

8 Thanks to Peter Godfrey-Smith for his particularly incisive comments on an earlier version of this paper.

9 I would like to thank the editors for inviting me to contribute to this collection, for amongst much else, it gives me the opportunity to place on the public record my gratitude to Steve, who was immensely generous to me both personally and academically very early in my professional career, when I needed it most. In being a recipient of Steve’s generosity, I am far from alone — he has consistently been an extraordinarily generous supporter both of his own students and of others early in their career. So, Steve, thanks.
References


