Evolution and Moral Realism

Ben Fraser and Kim Sterelny

Draft of November, 2013

1. Realism about Scientific and Normative Thought.

In the last decade, work on the evolution of moral cognition has greatly expanded (see (Richerson and Boyd 2001; Joyce 2006; Boehm 2012; Richerson and Henrich 2012; Chudek, Zhao et al. 2013; Sterelny forthcoming)). What do these evolutionary hypotheses tell us about the nature of normative judgements themselves? One response, and probably the most influential, has been to take these evolutionary hypotheses to undermine the idea that there are moral facts: moral judgments are shown to be false, or probably false, or unjustified (Mackie 1977; Ruse 1986; Joyce 2006; Street 2006). In particular, Michael Ruse, Richard Joyce and Sharon Street have made evolutionary error theory an influential way of connecting evolution to ethics[1]. The sceptical idea is that an evolutionary account of the origins and stability of moral thinking displaces an account of moral thinking as responding to moral facts. The idea of a moral fact is shown to be redundant, playing no role in the explanation of moral belief. At the same time, the argument shows that for moral thinking to play its regulative role in human social life, moralized and moralising agents must think of moral judgements as responses to moral facts, as only this explains their power to induce agents to act against their inclinations and interests.

One evolutionary analysis of religious belief is an influential model for this sceptical line of thought. On this analysis, religious commitment is adaptive, buying agents (or communities) the benefits of cooperation and social cohesion. But these benefits depend on agents' belief in the reality, power, and zeal of supernatural oversight of their actions (Wilson 2002; Bulbulia 2004b; Bulbulia 2004a). An evolutionary genealogy explains the persistence of religious belief, while showing that its adaptive benefits depend on religious commitments being taken to be truth tracking. At the same time, it debunks it, for the analysis shows that our being prone to religious belief

[1] These debunking views of moral cognition are premised on cognivist views of moral language and thought: moral claims are failed attempts to state facts: they are truth-apt without being true. It is worth noting that an eliminitivist view of moral thinking is not committed to a cognitivist analysis of moral language. Suppose moral judgement were (say) the expression of some distinctive form of disgust (or delight) and no more. We could still wonder whether we should continue to support the social institutions of moral response (for example, to invest in formal and informal moral education), and the genealogy of those institutions might well be relevant to that answer. So there is a version of the vindication problem that is not tied to the idea that moral language is designed to be fact-stating. We shall bracket off these issues here, but see Fraser (in preparation) for further exploration of these issues.

is not counterfactually sensitive to the existence of religious truths for those beliefs to track. We would believe in gods, whether gods were real or not. Likewise, we would have moral beliefs, whether or not there were moral facts.

An evolutionary genealogy of religion really does debunk religion. But religion is a poor model of moral thinking. We shall argue that human moral practices are a complex mosaic. Elements of that mosaic have different origins, respond to different selective forces; depend on different cognitive capacities; probably have different metanormative evaluations. After all, human moral practices include both fast, implicit, reflex-like online cognitive systems; slow, explicit, offline systems. They involve both individual cognitive mechanisms and collective institutions (for example, communities have a stock of stories and narratives that frame their moral education). It includes both the internalisation of individual values and the use of moral language to persuade others; we are both moralised and moralising. It would be no surprise if there were no unitary explanation and assessment of all these elements. But within the moral mosaic, we shall identify one important element in the genealogy of moral thinking, and argue that this strand of the genealogy of moral thinking supports reductive naturalism. Moral truths will turn out to be truths about human cooperation and the social practices that support cooperation. For moral thinking has evolved in part in response to these facts and to track these facts. So one function of moral thinking is to track a class of facts about human social environments, just as folk psychological thinking has in part evolved to track cognitive facts about human decision making.

The idea that connects moral thinking to the expansion of cooperation in the human lineage has two complementary aspects. First, it is important to an individual to be chosen as a partner by others: access to the profits of cooperation often depends on partner choice. Choice in turn is often dependent on being of good repute, and (often) the most reliable way of having a good reputation is to deserve it. It is worth being good to seem good. Recognising and internalising moral norms is typically individually beneficial through its payoff in reputation (Frank 1988; Noë 2001;

Baumard, Andre et al. 2013)[2]. Second, human social life long ago crossed a complexity threshold, and once it did so, problems of coordination, division of labour, access to property and products, rights and responsibilities in family organisation could no longer be solved on the fly, or settled on a case by case basis by individual interaction (Sterelny forthcoming). Default patterns of interaction became wired in as social expectations and then norms, as individuals came to take decisions and make plans on the assumption that those defaults would be respected, treating them as stable backgrounds; naturally resenting unpleasant surprises when faced by deviations from these expectations. The positive benefits of successful coordination with others, and the costs of violating other's expectations, gave individuals an incentive to internalise and conform to these defaults.

These gradually emerging regularities of social interaction and cooperation were not arbitrary: they reflected (no doubt imperfectly) the circumstances in which human societies work well, and how individuals can act effectively in these societies to mutual benefit. Given the benefits of cooperation in human social worlds, we have been selected to recognise and respond to these facts. So this adaptationist perspective on moral cognition suggests that normative thought and normative institutions are a response to selection in the hominin lineage for capacities that make stable, long-term, and spatially extended forms of cooperation and collaboration possible. On these views, there is positive feedback between moral thought and judgement and the distinctive forms of human social life. The conditions of human sociality selected for, and continue to select for, normative response, and the emergence of norms allowed those distinctive forms of social life to stabilise and expand, further selecting for our capacities to make normative judgements.

A natural notion of moral truth falls out of the picture that moral belief evolved (in part) is to recognise, respond to, promote and expand the practices that make stable cooperation possible. For there are objective facts about the conditions which make cooperation profitable, and about the individual capacities and social environments which make those profits more or less difficult to realise. For example, evolutionary game theory has shown the importance of group size, interaction frequency, and

[2] This signalling or advertising function of moral response can be seen as a reason to be sceptical about truth-tracking views of moral cognition: see (Fraser 2012; Fraser 2013)

cheap and reliable information (Bowles and Gintis 2011). There are also objective facts about the practices and norms which would promote stable cooperation within the group. Evolutionary game theory is helpful here too, since its analysis often shows that distinct equilibria — different stabilised patterns in behaviour that become customs and norms — differ in their capacity to deliver cooperation profits. No doubt there is no single set of optimal norms: the best normative packages for a group will depend on its size, heterogeneity, and way of life. No doubt there are trade-offs between the size of the cooperation profit and its distribution. But despite these complications, a natural notion of moral truth seems to emerge from the idea that normative thought has evolved to mediate stable cooperation. The ideal norms are robust decision heuristics, in that they satsfice over a wide range of agent choice points, typically providing the agent with a decent outcome, in part by giving others incentives to continue to treat the agent as a social partner in good standing. They are robust as well in being good heuristics in a range of normative environments; their positive effects do not depend on a very specific set of other normative beliefs. The moral truths are those maxims which are members of all or most near-optimal normative packages; sets of norms that if adopted, would help generate high levels of appropriately distributed and hence stable cooperation profits.

On this view of moral thinking, as with other neo-conventionalist accounts, moral thinking emerges as a version of prudence (in this respect, our views are similar to those of (Gauthier 1987). In general, agents have an individual stake in supporting effective yet fair cooperative practices. We might prefer unfair solutions if we were to be part of the elite, but fairness typically satisfices. A fair social world might not be our first choice, but it is certainly not bottom of the list, and so it is rational to choose fair norms from behind an evolutionary veil of ignorance[3]. We are moral only because it was and is in our interests to be moral. But our evolved dispositions make us *genuinely moral*. Moral response is not voluntary, not conditional on individual decision or calculation at particular choice points. To borrow a term from Daniel Dennett, our commitment to moral policies is ballistic, rather than being re-evaluated on a moment by moment, case by case basis (Dennett 1995). We do not decide on a

[3] Though there is scope for considerable variation in what counts as far norms: see (Baumard, Andre et al. 2013). That is fine by us, as we do not think there is a uniquely package of optimal norms. Many different sets will make good enough trade-offs between fairness and incentivising individual effort for cooperative social life to be stable.

case by case basis to feel moral emotions or to make moral judgements. Sometimes, then, thinking and acting morally will not be in an agent's interests. But because such cases are atypical, we have been selected to genuinely endorse moral views, even though it would sometimes be in our interests to ignore them. Moreover, as with any form of naturalism, on this view of moral thought, the motivational force of any moral claim is extrinsic to its content[4]. In a sense, there is objective, independent moral authority, but only in these sense that the cooperation phenomena do not depend on our moral labels. There power to motivate us is contingent, but nonetheless deriving from developmentally and evolutionarily deep and relatively inflexible features of typical human personalities.

In this paper, then, we have three targets. First: we aim to locate error theory and reductive naturalism within the broader context of the relationship between science and the folk frameworks for thinking about the world. There are many domains of folk thought in which folk commitments do not seem to fit naturally with a developing scientific consensus about the world and our place in it. We intend to extract and exploit conceptual tools developed for these other cases, and use them in evaluating evolutionary error theory. Second, we shall suggest that these other cases undermine the dichotomy between reduction and elimination; between error theory and vindication. Suppose, for example, that the folk really do think of conscious states as ones on which we act, and which are known directly and incorrigibly. Suppose further that the cognitive sciences tell us that no action guiding states are perfectly detectable. It would not follow that there were no conscious states. Folk conceptual frameworks can be imperfect, yet still latch onto and partially describe important phenomena in our environments, and guide action with respect to those phenomena.

Third, we argue that the mosaic character and complex genealogy of moral thinking and practice are important. Very likely, moral thought and judgement in part evolved to facilitate mutually profitable social interaction by tracking and responding to

[4] In our view, this is true of any theory that treats moral claims as anything more than a disguised version of *actual* desires. Suppose, for example, we view moral claims as the preferences of ideal observers; or the preferences that agents would converge on, under conditions of impartiality and objectivity. Once the sophisticated noncognitivist defends some form of idealising analysis, they too must treat the actual motivating force of endorsing a moral claim as extrinsic to the judgment itself. It makes perfect sense to wonder why we should give want what our ideal self would prefer.

roadblocks that limit cooperative profits. But that is only one factor in the matrix of selection through which human moralising emerged. In Lewis-Skyrms signalling systems, signalling emerges when one agent, "the receiver" can act, but lacks information about the environment that only a second agent, "the sender" has access to. If the receiver acts to their mutual benefit, signalling emerges. These signals both track variable states of the environment, and guide adaptive response, and this is the core from which indicative, truth-apt language was built (Lewis 1969; Skyrms 2010). But signalling systems can emerge as pure coordination devices, when agents have an interest in mutually adjusted interaction, as in dance and many games (Godfrey-Smith 2012). In such cases, the benefits of coordinated interaction do not depend on the coordinating signals' match to some independent, variable feature of the world, and so these signals do not have any world-tracking function. Linguistic conventions about (for example) word use are coordinating devices like this, and we shall argue that moral norms in part play this pure coordinating role too; one that does not depend on them tracking independent features of the social world. They may well have other roles as well: as aspects of sexual display; as devices for both marking and deepening ingroup-outgroup distinctions.

This complex genealogy is relevant for two reason. First, to the extent that these non-tracking functions are important, we would expect genetic and cultural selection to be less effective filters, less effective in predisposing us to norms that do actually promote profitable, stable cooperation. Second, to the extent that these non-tracking functions are important, our view of the evolution of moral thinking is a less persuasive consideration in favour of a cognitivist view of moral thought and talk. If moral thinking evolved as a tracking device, selected to track and respond to cooperation pitfalls, then the apparently truth-apt character of moral thought and talk would reflect its functional role. The less it evolved as a tracking device, the less apparent form reflects role, and the more plausible non-cognitivist options become.

The road ahead is as follows. The next two sections are on reductive naturalist hypotheses in general. When should we discard folk frameworks; when should we regard them as largely vindicated by our best science? These sections are about the interpretation and assessment of folk frameworks, and we conclude with an

intermediate case, folk astronomy. We think folk astronomy is important, because it supports adaptive action quite flexibly, despite astronomical belief being embedded in a seriously mistaken set of general beliefs. In our view, folk astronomy is a good model for normative belief. In section 4, we make that case. In particular, we argue that the appeal to moral facts does explanatory work. Error theorists (and not only error theorists) argue that to explain human social life, we must appeal to human moral opinion, but we need make no mention of moral facts. In Section 4, we respond to the argument that moral facts are epiphenomenal; the argument that moral opinions are causally important but moral facts explain nothing. In section 5 we make a positive case for the explanatory importance of true moral beliefs; we show that in some important ways, they are maps by which we steer. We then summarise and conclude.

2. The Folk and Science

One of the great projects of contemporary philosophy is to explore and identify the relations between two apparently different ways of thinking about humans and their place in the world. We develop one set of cognitive tools from our socialisation as members of our communities: we develop folk understandings of the physical world, the biological world, human agency and so on. This is Sellar's "manifest image", though we should of course speak of manifest images, as there is no single folk framework for thinking about the world: it has varied across time and culture. The other conceptualisation has developed within the natural and social sciences over the last few centuries: the "scientific image" of humans and of the world in which they act. The two conceptualisations are not obviously compatible: how does (say) the view of human agents as self-aware, rational decision-makers fit with the view of humans as a modified great ape? Can we show that, perhaps despite appearances, the folk conception is compatible with the one developed from science? If not, how should we respond?

One major move in this philosophical space is reductive naturalism. The key strategy is to co-opt an idea developed in understanding the relationship between sciences, and use it to understand the relationship between folk thought and scientific thought.

Within science itself, reduction is the claim that the facts in one domain — a reduced domain — are less explanatorily fundamental than the facts in a reducing domain. For example, facts about the temperature of macroscopic objects are less fundamental than facts about mean kinetic energy of the molecules that compose that object. Facts about inheritance patterns between parents and offspring in sexually reproducing population — the facts systematised and predicted in classical genetics — are less fundamental than facts about the sequences of DNA base pairs in the haploid gametes transmitted across generations, the gametes that fuse to form a new individual. Facts about the stability of a species over time are less fundamental than the facts about the flow of genes in that species' gene pool, and their constrained flow outside that gene pool. The reducing facts explain the reduced facts, but not vice versa. The most plausible cases within the sciences (perhaps the only plausible cases within the sciences) are relations of composition. The character, distribution, and interaction of the parts of a system explain the behaviour of the system as a whole. Thus a gas is made of molecules, and it is their character and interaction that explain the properties of the gas.

Reductive naturalists extend this idea to folk kinds. Famously, at the dawn of this project, Jack Smart suggested that facts about conscious experiences of pleasure and pain were explained by, and reduced to, facts about human neurophysical organisation and activity. Reductive naturalists point out that the reductive relationship between domains — even when it involves kinds recognised by the folk — is not *in itself* a piece of folk wisdom. For these reductive relations between domains are discovered empirically, they are not a priori or conceptual truths. One of the stock examples is the identification of water as H_2O; now widely known as a chemical factoid, and an example of compositional reduction, but once a major discovery of scientific chemistry. This is no surprise on a system-component model of reduction. The folk will often be acquainted with, and have reliable information about, complex macroscopic systems — like organisms or agents — but be without systematic access to information about their internal components and their organisation. So naturally humans can develop a concept of water, and know lots of truths about water, without knowing that it is nothing but a configuration of oxygen and hydrogen atoms.

As we have just noted, the water=H_2O example depends in part on the idea that we can have epistemic access to a system without thereby having epistemic access to its components and organisation. A second model, exemplified by the notorious identity of the morning and the evening star, depends just on this idea of distinct routes of epistemic access. The thought here is that we can have epistemic access to the same individual or kind through two different routes; and form a variety of true judgements about a kind or its instances without realising that there is only a single kind in play. Early versions of materialist theories of mind — early forms of functionalism and Smart's "topic neutral" analysis of mental concepts — had to explain why the identity of mind and brain (if they were indeed identical) was not obvious to all of us. Their response was initially ambiguous between a two-routes model of epistemic access and a system-component model. But later versions of functionalism — the functional decomposition models of Dennett, Lycan and Stich — are clearly system-component models. The point, though, is that any form of reductive naturalism targeted at folk kinds needs some account of how the reductive identity can be true, without its being known to be true, despite the fact that folk agents have plenty of information about the reduced domain, and sometimes even the reducing domain.

The folk know about physical interaction, animals and plants; they know that other agents have minds. But they also know, or seem to know, about norms and values. We think of actions as cruel or kind; generous or stingy; required or forbidden. We think of some people as admirable, and others as arseholes. To put it floridly, typical humans take themselves to live in a normative world not just a physical world. How does this aspect of the manifest image relate to the scientific image? There is a version of the reductive naturalist project, known as "Cornell Realism" (David Brink, Richard Boyd) that extends that project to normative phenomena. Norms turn out to be natural facts.

The Cornell realists take water=H_20 as their paradigm for thinking about the relationship between natural and normative facts, because this paradigm blocks the open question argument. It shows that a reductive identity can be true without its truth being apparent to any cognitively and linguistically competent member of a community. But in other respects, it is a misleading model. For one thing, there is no

composition relationship between normative kinds and any plausible set of base properties. The reduction base is a set of facts about agents, their interests, the social systems of which they are a part, and the deep history of those social systems. Dan Dennett's picture of the intentional stance offers a better model of the relationship between the normative and the natural. The cognitive and neural organisation of an agent sharply constrains the belief-desire profiles we can attribute to that agent (Dennett 1991b). So there is a very important relationship between belief-desire psychology and cognitive psychology and cognitive neuroscience, and that is true even though our practices of interpreting one another change our cognitive organisation, perhaps very profoundly (Mameli 2001; Ross 2006; Zawidzki 2013). But while these constraints are important, they do not uniquely specify an intentional profile. Dennett argues that no agent's actual behavioural dispositions will perfectly match any intentional profile; such profiles always idealise behavioural patterns to some extent. Profiles can be distinct, but equally legitimate, because they make different trade-offs between simplicity and accuracy. Moreover, while the cognitive organisation of an agent explains their behavioural dispositions and hence their intentional profile, specific beliefs and desires do not routinely map onto specific elements of an agent's cognitive organisation (see especially (Dennett 1991b).

The relationship between intentional and cognitive psychology is a better model for the realist, because it is not a system-component view of the relationship between domains (beliefs, for example, are not composed of specific neurocognitive structures). It does not commit us to the view that there is a unique set of moral truths fixed by the reduction base, and nor does it commit us to the view that there is an element by element reduction of normative predicates to natural predicates. That is important. On the hypothesis we have been considering, the reduction base is a set of facts about human communities (including ancient ones): facts about profitable forms of cooperation, about social arrangements and cognitive dispositions positively and negatively relevant to the stable exploitation of those opportunities. These complex social environments selected for human recognition of, and response to, maxims of social interaction which in general improved human access to these benefits. But it is most unlikely that there is an element to element mapping between the norms in

adaptive packages, and opportunities and barriers to cooperation. Norms are typically relevant to many action choices in many contexts.

3. Reduction, Vindication and Error

The reductive project, when carried through successfully, is intended to vindicate the folk conception of the world. A theory of free will, for example, might identify free action with fallible but still coherent and informationally sensitive decision making, and show that on important occasions humans make decisions of that kind. Such a theory would vindicate the idea that human agents sometimes act freely. Contrast that with a sceptical theory that emphasised our ignorance of our own motivational structure; our cognitive biases, and the sensitivity of action and judgement to clearly irrelevant contextual factors. A theory of this kind would be best seen as explaining the illusion of free will. But this free will example raises a methodological challenge to the project of naturalistic mapping. How should we distinguish genuine from ersatz mapping? Thus it's often claimed that the folk are committed to the idea that there is real free will, real autonomy, not merely the (approximate, fallible) capacity to act with appropriate informational sensitivity on the basis of a stable, reflectively assessed preference order. Likewise, Dennett's Consciousness Explained often met with the charge that Dennett was explaining consciousness away, not explaining it (Dennett 1991a). So what counts as vindicating the folk idea that conscious thoughts are inner episodes, by showing that the folk were onto something real, according to our best science, and what counts as explaining away the folk illusion that there are such thoughts?

The ersatz problem makes it natural to link the project of naturalistic mapping to one of philosophical anthropology. The idea is to take a domain of folk opinion (in our case, normative thought and talk) and attempt to systematise, in the best possible way, that opinion. This project proceeds by a mix of methods. Ideally, a systematisation of (say) the folk concept of consciousness will capture both the folk's intuitive judgment of specific cases — I am conscious right now as I read this paper — and the folk's general principles about consciousness: for example, that it is psychological state, but not one agents are in all the time; whether you are conscious of a particular event is

relevant to whether you enjoy it; adult humans engaged in ordinary mundane activity are conscious; rocks and corpses are not conscious, and so on. The "Canberra Plan" is a particularly well-developed and theoretically well-motivated version of philosophical anthropology (Jackson 1997). The Canberra Plan is alert to the fact that we should expect there to be noise: we should not expect the folk to be completely unanimous; we should expect there to be marginal or debatable examples of folk maxims (in this case: perhaps whether consciousness comes in degrees); we should expect some failure of fit between judgments of particular cases and the systematisation of folk principles. Something like reflective equilibrium will play a role in the reconstruction of the folk conception of a domain.

Moreover, the maxims need not all be of equal importance. Some will be more central to the role the folk concept plays in organising action. Perhaps in the case of consciousness, maxims about the relationship between conscious experience and affective valence are more central than those about noninferential knowledge of conscious states. It is also true that folk maxims, especially the general principles, are rarely explicit features of folk frameworks. So part of the project is making implicit commitments explicit, typically by reflection on intuitive judgements about particular cases; a procedure that leaves plenty of room for uncertainties. So there are problems and complications, but once we have recognised the noise, and hence identified the core folk commitments about (say) consciousness, we have, in effect, constructed an implicit definition of consciousness: consciousness is that unique state X that satisfies the following conditions: X is mental state; awake, normally acting adult humans are in X; rocks are never in X; and so on, for all the clear, core, uncontroversial features of the folk's view of what it is to be conscious.

Once we have done that, we have constructed a potential bridge between putative folk kinds and our best science. For we can then ask, from the perspective of our best science, whether there is any unique kind that satisfies the conditions specified in the implicit definition. Philosophical anthropology might show that stable, self-reflective rational decision making is necessary and sufficient for free action. It would then be the task of cognitive and social psychology to determine whether human decision making regularly (or ever) satisfied these conditions and, in particular, whether it does

so in those cases regarded as paradigmatic of free action. If not, we should be error theorists about free choice. To take an actual and clear example, we should be error theorists about the racial views that became popular in the UK in the nineteenth century: the view that there were identifiable, perceptible differences between northern and southern Europeans; these were probably originally caused by climatic and geographic differences (with the northern peoples rising to the challenge); that the perceptible differences indexed morally important differences in intelligence and personality; that the differences were now innate; individuals of mixed descent tended to have the character of the lower type (alternately, the virtues of neither); the differences explained the differences in economic wealth and political power between the north and the south. On this understanding of what a racial type is, there are no racial types.

Thus some folk frameworks have rightly been discarded, but one of the strengths of the Canberra Plan is that it very naturally recognises the fact that there are many cases intermediate between vindication and error theory. There may be a unique state that satisfies some but not all of the clauses in an implicit definition of free will or conscious thought; there might be a state that satisfies all or most of the conditions, so that some human actions are free, but it turns out that those cases considered to be paradigms of free action are not free. Suppose for example, that agents often make very good decisions when they make fast, on the spot judgements in situations in which they are very experienced, showing just the right sensitivities to subtle differences in circumstances. But when they attempt to make good decisions through explicit, slow, careful self-conscious reasoning, they are especially prone to framing effects and irrelevant contextual cues. The idea that we make free choices would then be neither vindicated nor debunked.

However, despite its capacity to recognise intermediate cases, the Canberra Plan seems to over-count failed folk frameworks. Consider thought and talk about the stars and planets in ancient world (the thought systematised and quantified in the Ptolemaic astronomy)[5]. A systematisation of Mediterranean astronomical thought of AD 200

[5] Thus this example is not strictly speaking a folk framework, since it includes elements that are produced by cultural elites — like calendars and almanacs — which are then absorbed into the general

might suggest that we should be error-theorist about pre-Galilean astronomy. Almost all of the general beliefs were mistaken, as were some of the particular identifications (the moon and sun are not planets; the earth is). Yet that does not seem right, for agents in the ancient world were able to use astronomical information adaptively in navigation and to tell the daily and seasonal time. Of course there is wriggle-room. For example, the Canberra Planer can insist that the maxims with the heaviest weight are one like "You can only see the stars at night" or "the stars do not seem to move in their relative positions but the planets do" or "Mars looks reddish when it is brightest". But this does seem to shift away from the idea that the agents in question had an (implicit, noisy) coherent conception of the night sky, that we can systematise and then attempt to map onto our best scientific conception. For it does not seem plausible that the agents themselves would have regarded these banal maxims of perceptual observation as their most central astronomical beliefs; especially once astrology took hold of both the lay and the educated mind.

We think that this example both shows the importance of know-how, skill, and suggests that skill is not just a special case of propositional knowledge. Sky-watchers of the ancient world had a complex of explicit beliefs about what they could see, but they also had a complex of discriminative capacities. They could identify and re-identify specific celestial objects and configurations, and those discriminative capacities supported adaptive action: direction finding, for example. Folk cognitive frameworks, on this view, are not just systems of propositional representation, and these frameworks can enable agents to register features of their environment, and guide response to them in ways that partially screen-off mistaken belief, sometimes even when those mistakes are quite fundamental. The folk can sometimes respond in quite nuanced ways without that response being routed through a conceptualisation of the phenomenon in question (presumably non-human animals typically manage their environment this way). In assessing folk belief systems we need to identify the features of an environment to which a folk cognitive practice is a response. Folk frameworks can be responsive to real phenomena, and guide action appropriate to

---

practice of the community. We do not think this complication affects the general point the example illustrates.

those phenomena without accurate conceptualisations of those phenomena[6]. The ancient world registered and responded to features of their celestial environment, and this guided navigation, calendar construction and time keeping. We think something similar is true of moral response; especially automatic, reactive moral responses. These depend on implicit generalisation from exemplars, rather than on explicit principles of moral reasoning. Moral cognition is partially know-how, it is not just a structure of propositions (Stich 1993; Churchland 1996; Sterelny 2010).

Discriminative capacities and the banal but true beliefs that they support help distinguish ancient astronomy from other apparently mixed cases. For example, taboos often support adaptive behaviour (Harris 1985), but in rigid and limited ways. In certain Amazonian tribes, fish-eating fish are a taboo food for pregnant women (Begossi *et al*. 2004). As it turns out, these fish contain high concentrations of toxins in virtue of being near the top of their food chain. So the tribesfolk are acting adaptively in identifying the fish as having this apparently spooky property and so avoiding it, but one might not think this much of a vindication. Suppose, though, that in addition to avoiding these fish, these agents had a way of identifying the toxin wherever it was found (suppose it to have a distinctive colour when baked), and always avoid it. The practice would still be embedded, as with ancient astronomy, in a deeply mistaken theoretical framework, but with the support of these discriminating capacities, identifying taboo substances would support adaptive action quite flexibility (it would be a fuel for success). The taboo case might be like some of the more successful elements of ancient and folk medicine. For some diseases and injuries have long been identified, and to some extent effectively treated, despite these practices being embedded in very mistaken theory. Malaria became such a case. By the seventeenth and eighteenth century, European physicians were aware of the connection between malaria and exposure to wetlands, and the use of quinine was becoming standard. But they had no clue about the aetiology of the disease or the

[6] This line of thought had its origins in the causal theory of reference: in the idea that agents could use names to refer to individuals about whom they had confused or mistaken ideas, and that they could refer to natural kind, without any idea of the key features all instances of the kind had in common. It is central to this line of thought that the folk capacity to think and talk about kinds and individuals does not rest on a correct implicit theory of those individuals. But there has been an ongoing debate on whether there are less demanding informational preconditions: on the idea that speakers and thinkers have to have some informational connection to the objects of thought and talk (Devitt and Sterelny 1999).

cure: ‘malaria’ derives from ‘bad air’, and it was thought that the disease was caused by vapours rising from swamps; equally, quinine had been introduced as a lucky guess; South America Indians used it to reduce shivering when they were very cold (Rocco 2000). Again, we shall suggest that moral cognition shares some features of this case, in that it supports a quite flexible responses, rather than an appropriate response only in a single stereotyped situation, as in the actual fish-taboo above.

So malaria is another mixed case, but to repeat the lessons of early racial theory, not all cases are mixed. To recycle a clichéd example, seventeenth century witch theory was a folk framework that deserved elimination. Even if those persecuted were an identifiable subgroup (friendless, isolated, socially deviant) rather than an ad hoc collection of the unlucky, discrimination did not leverage adaptive behaviour, even by the lights of the witch-burners. It did not prevent crop failures or other misfortunes. So nothing in the world remotely corresponds to the witch-identifying maxims; nor did witch representations leverage adaptive behaviour.

So vindication is possible, and so is elimination. But we think that the most important upshot of this discussion is that it is a mistake to frame the discussion of folk ontologies as a choice between reduction and elimination. In many cases, that framework is misleading. Many case, perhaps most cases, will involve some mix of vindication and elimination; some mix of mere causally grounded response to phenomena in the world; partially correct conceptualisation and description of those phenomena; and some capacity to support effective action through tracking and conceptualisation. In particular, the ancient astronomy example shows that folk conceptual systems can systematise responses to phenomena in the world in ways that leverage adaptive behaviour, even though those conceptual systems mis-describe their targets in genuinely important ways. Despite the errors in premodern astronomy, premodern astronomical beliefs leveraged adaptive actions from those own agents’ own perspective regularly and systematically. Premodern beliefs about witches did not, since burning witches did not stop crop failure, plague and other local disasters, or even expel the devil. From the perspective of the witch-finder’s own ends, witch killing was ineffectual. So witch lore did not give agents theoretical leverage over the nature of the world, and nor did it give them practical leverage in making things

happen. Ancient astronomy gave a little of the first, and quite a lot of the second. Ancient astronomy, then, is a mixed case.

4. Moral Facts and Moral Opinions

Prima facie, we would expect folk moral theories to be at best a mixed case, too. We noted in the introduction its mixed genealogy. While moral thinking evolved to track the social environment, it did not evolve only as a tracking device. Moreover, hidden-hand mechanisms are far from perfect in producing optimal adaptations to heterogeneous and fast-changing environments (Sterelny 2007). The wide variation in moral opinion seems to confirm this pessimistic expectation, showing that if moral thinking tracks moral facts, it cannot be doing so very efficiently. Perhaps some variation is sensible adjustment to different local circumstances. But some of it is surely real, and where it is real, not everyone can be more or less right. One source of error is that, as with ancient astronomy, in many cultures moral thinking keeps bad company, being entwined with bizarre religious misconceptions, local origin myths, dubious politics, crackpot notions of purity and health. In addition, there is often at least some self-serving influence of elites on local moral opinion. So no adaptationist conception of the evolution of moral thinking will deliver a full, clean vindication of diverse moral opinion. Indeed, we expect the moral case to be intermediate in a variety of respects. First: our moral practices are a mosaic; some elements may turn out to be vindicated, others revised, others discarded. Second: as we have noted, moral judgements functions to signal and to bond, not just to track; vindication is only in question with respect to tracking. Third, as we shall now explain, tracking is only partially successful.

As we see it, to even partially vindicate folk moral theory, the evolutionary realist must meet two challenges. First: error theorists have argued that the appeal to moral facts or moral truth is redundant: we can explain human moral thought, and the influence of thought on action, without appeal to moral facts. Second, the evolutionary realist needs to develop a positive case analogous to the one noted for ancient astronomy. In thinking about astronomy, we saw that despite theoretical misconceptions, many folk astronomical beliefs were true (even though they were

pretty mundane observational beliefs) and that the cognitive network of astronomical beliefs and the perceptual capacities that supported them powered adaptive action quite flexibly, and over a range of contexts. Folk representation of their celestial environment was, to some extent, a "fuel for success". Can we show the same about folk moral thinking?

We begin with the issue of redundancy. A core idea in error theory, including evolutionary error theory, is the claim that the appeal of moral facts is redundant. In explaining human social life we must appeal to agents' moral opinions. But we never have to appeal to the supposed moral facts that these opinions track. Evolutionary error theorists argue that moral facts, if they exist, are epiphenomenal, as they play no essential role in explaining moral response (the argument derives from (Harman 1977)). Philip Kitcher has recently presented a version of this argument, recycling a stock example: moral anger at the sight of a dog being wantonly and severely tortured. He argues that while moral anger is immediate and involuntary, we should not think of such responses as the quasi-perceptual detection of an objective moral property. Rather, it is a result of social learning: we learn to respond emotionally to a certain class of situations. He recognises that these immediate responses are phenomenologically similar to another class of learned responses: trained expert response to the otherwise cryptic perceptual outputs of experimental apparatus. But he thinks there is a critical difference between the two cases. Think of a scientist reading bubble chamber tracks or gene sequences in a gel. In these cases, the expert practitioner has a learned sensitivity to subtle perceptual detail that otherwise would not be salient, to the patterns these details form, and to what those patterns say about the causal processes that produce them. After a long period of scientific socialisation, these interpretative responses are as immediate and involuntary as our response to the sight of a dog being wantonly tortured. But (the argument runs) despite the phenomenological similarity between laboratory expert and folk response to outrageous behaviour, scientific expertise really is a quasi-perceptual recognition of objective phenomena. A detailed history of the lab skills involved in interpreting cloud chamber photographs essentially involves facts about the subatomic world; their interactions in bubble chambers; the macroscopic patterns they generate, and the perceptual discrimination of those patterns.

We think Kitcher's diagnosis of scientific expertise is right. So if moral response is quasi-perceptual, moral facts (as our evolutionary reductive naturalist construes moral facts) must play a crucial role in the origins of these responses (and, presumably, to less automatic forms of moral thought). At this point in the argument, moral diversity becomes salient. As we remarked above, some diversity of views may well reflect genuine differences in the social environment. For example, in many earlier societies punishment for norm violation often seems to us to have been extraordinarily severe, even where we agree that the norms themselves were warranted. But this might reflect a genuine difference in two social worlds; perhaps a smaller social surplus, so those could not afford to waste resources on people in gaol, or perhaps harsh punishment compensated for less reliable detection of violations. But sometimes this explanation looks implausible. Remember, for example, that torturing animals (and people) was once popular entertainment. Badger-baiting (for example) was popular in England in the eighteenth and nineteenth centuries (and apparently is not extinct). The contrast with clearly perceptual cases is clear. There is no diversity in the view that badgers are animals, bigger than rats and smaller than pigs. But there is diversity on whether inflicting pain and death on them is disgusting or fun.

That diversity of views challenges the idea that there is an opinion-independent fact about whether badger-baiting is wrong. For the diversity of scientific thought has often been seen as trouble for scientific realism, and a parallel argument against moral realism seems even stronger. Suppose we construe scientific realism as the following three claims: (i) there are observer-independent facts about the natural world; (ii) The aim of science is to identify and represent those facts (presumably, in as compact and as systematic a way as possible); (iii) the sciences have had significant and growing success in carrying out those aims. The diversity of scientific ideas is a threat to clause (iii). The "pessimistic meta-induction on the history of science" reads the history of science as the history of fundamental change of opinion on the nature of physical reality. According to this view of the history of science, by our current lights, previous generations made important misidentifications of the physical world, despite their confidence in their own views. But our epistemic situation is no better than their situation, and so we should expect future generations to discard our current

conceptions. If the historical fluidity of scientific thought undermines scientific realism, surely the historical and cultural fluidity of normative principles undermines moral realism. A pessimistic meta-induction on the historical and cultural variation in moral opinion concludes that our response to cruelty is merely our response to the historical accidents of our life and times; in other times and places, people have learned to respond differently as a result of their own accidents of circumstance. Variation undermines the idea that moral response tracks, even fallibly, moral facts.

Scientific realists respond to the meta-induction by showing that it both exaggerates the degree of change in scientific view, and understates the epistemic difference between current and early science. Those are our options here, too. Can we show that despite appearances, there is relatively little difference in moral views; alternatively, can we make a claim to moral expertise, to a privileged standpoint with respect to eighteenth century badger-baiters? There are those that deny that there is much genuine variation in moral judgement[7], but we shall instead develop the second option. While every moral belief is the result of social learning, not all social pathways to belief are equal.

All moral learning involves an interaction between our systems of social emotion, individual trial and error learning (as children explore in and negotiate their social

[7] In particular, moral nativists think that while the diversity of moral opinion is real, it masks cross-cultural similarity at a more abstract level. For example, a specific wrong to another agent is more culpable if it is the result of deliberate act (especially if it's a direct and immediate effect of a deliberate act) than it would be, if the harm were the consequence of a failure to act. Likewise, nativists argue that there is cross-cultural sensitivity to the distinction between foreseeing an evil consequence of an action, and intending an action to have that evil consequence. Thus it is generally regarded as permissible to avoid a greater evil at the cost of a lesser one, when that lesser one is a side-effect of an action. But not when the lesser evil is the direct immediate consequences of an act (Hauser 2006; Mikhail 2007).

We might try the idea that these principles are genuine universals of moral reasoning because recognition of, and respect for, these maxims was and is adaptive for individuals and for the communities of which they are a part. The suggestion is not ad hoc. Utilitarians have regularly, and quite plausibly, argued that intuitions of this kind make consequentialist sense. There are good social engineering reasons to ensure that prohibitions not to harm have more teeth than requirements to help, and given human fallibility, we should generally deter deliberately and directly inflicting harm on others, even when the agent genuinely believes this is the only way of avoiding worse. If these principles are excellent heuristic guides to wise choice, and if our systems of moral thought are evolved innate adaptations designed to allow us to benefit from life in a cooperative world, these are just the principles we would expect selection to wire into our heads. According to this argument, core features of moral judgement are not the result of socialisation at all, let alone arbitrary, historical accidents of socialisation. But we put little weight on this response, for the empirical case for moral nativism is not strong (Sterelny 2010).

space) and the moral opinions of their community. These community opinions are expressed tacitly in their actions and interactions with one another; less tacitly in their customs and institutions; explicitly in their normative vocabulary, explicit moral maxims and narrative life (for a more detailed exposition of this view of moral learning, see (Sterelny 2010; Sterelny 2012)). However, though all social norms are acquired by some form of social learning, not all learning pathways are equal. The crucial constraint, then, on a naturalising theory of moral opinion is that it reveals a systematic difference between the history of error and of truth[8]. According to the evolutionary moral naturalist, the natural history of true and partially true moral views (as the evolutionary naturalist specifies moral truth) is different from the natural history of error and prejudice, in the immediate psychological history of individuals, or in the social context of social learning, or both.

Given the general nature of human social learning, and even given the mixed genealogy of moral practice and the sources of error we noted at the beginning of this session, we suggest that there are three ways that agents become aware, with some reliability, of the opportunities and challenges of human cooperation, and come to endorse norms that improve access to the profits of cooperation: (i) learning guided by prosocial emotions; (ii) vicarious trial and error learning in heterogeneous environments; (iii) cultural group selection. We begin with prosocial emotions. Jesse Prinz and Shaun Nicholls argue that while norms are learned socially, some norms are especially salient. There is a particular learning route that goes via our recognition of emotional response: cases where our acts affect other agents about whom we care, and we notice both their emotional responses to our actions, and our emotional responses to their responses. Thus generosity to others is readily reinforced through a loop in which their positive response induces your own positive response through emotional contagion; a response which you yourself notice. This does not guarantee the acquisition of norms of sharing (nor norms of harm avoidance, in the negative case). But it does make the phenomena that fall under those norms salient (Nichols 2004; Prinz 2007). Salience is no guarantee of truth. But if, as the evolutionary

[8] Ideally, we would also expect that natural history to guide methodological improvement in normative thinking: to reveal ways in which, individually or collectively, we can make moral thinking more reliable. For the natural history of scientific reasoning is a guide to the future: in unmasking biases (for example, palaeoanthropological narratives blind to female agency), we add to the checklist used to scrutinise new theories.

naturalist supposes, our species has had a long history of biological and cultural selection in favour of cooperation-supporting emotional responses, patterns of behaviour we find emotionally repugnant are likely to be instances of behaviours that would be forbidden by socially efficient norms; those we find appealing are usually instances of behaviours that would be endorsed by socially efficient norms.

Second, many contemporary societies are normatively heterogeneous, composed of cross-cutting groups with competing norms and agendas. In these heterogeneous contexts, agents have some ability to treat each other as natural experiments. In interacting with others who embrace different normative packages, we have some opportunity to see how their lives go: do they live in networks of support and mutual aid; are they regularly exploited by freeloading neighbours; are their lives blighted by moral feuds and the enmity of former friends? While a full-blown evolutionary perspective on the origins and stabilisation of moral cognition is not part of folk wisdom anywhere, the idea that norms have a social role that promotes fair interaction may well be, and in these mixed learning environments, that awareness may play some role in the norms agents accept and internalise. After all, moral education quite often proceed by noting the effects of norms, and hence norm violations, on cooperative lives ("What if everyone did that dear?").

In this respect, moral norms are very different from the religious norms we discussed in section 1: the social role of moral norms can be transparent to end-users without that knowledge eroding their role. While evolutionary models of the emergence of norms do not presuppose that agents understand the role norms play in their lives, they do not presuppose that they have no insight into this role. Indeed, because normative facts are mundane facts, ordinary agents have access to many of them, and so folk thought and argument about norms is not futile. On this evolutionary naturalist picture, there is nothing mysterious about moral epistemology. That would be different if, say, the truth-makers for normative claims were historical facts about the Pleistocene. Moral knowledge is not knowledge of mysterious or inaccessible facts.

Third, in the past, communities were smaller and more internally uniform. Some of these communities did well; others less well. Arguably, one causally relevant factor was the extent to which their normative lives stabilised and enhanced local cooperation. Cultural group selection will favour systems of moral norms that are relatively efficient means to the ends of social peace, regulation of conflict, and the restraint of selfish or destructive impulses (Boyd and Richerson 1990; Bowles and Gintis 2011; Chudek, Zhao et al. 2013). Evolutionary naturalists, then, are committed to theories of the evolution of norms that see this as an ongoing process of gene-culture coevolution. But this is independently plausible: human cognitive evolution did not stop in mid-Pleistocene Africa.

So the outcome of norm learning is not just luck, not just the outcome of local accidents. These mechanisms can be overridden by other processes, and even when they guide norm acquisition, they are by no means guaranteed to guide agents to true norms, to one of the optimal packages. But given a social and physical environment, and a set of interacting agents with their opinions and motives, there will be facts about whether their current norms are efficient means to stable and profitable cooperation. And to the extent that norms that do support cooperative interactions establish in a culture, it is typically not just by lucky accident. There is some tendency for better norms to be found, though, this process is noisy, imperfect and dependent on deep evolutionary histories, not just intelligent individual learning.

If this is right, actual systems of moral opinion will be a mixed bag. The naturalist project is to show that the elements in this bag tend to have rather different cultural histories, and depend on different social learning processes. Some will be unfortunate historical legacies (lingering prejudices of various kinds). Some will levers for exploitation and injustice that exist because of imbalances of power and wealth. Some will indeed be the result of selective filtering, but not for tracking and responding to levers of cooperation. But some actual maxims will be true and their truth will have played an important role in their becoming widely endorsed. For example: it is surely likely that the maxim: do not be cruel, or the maxim: do not inflict severe pain for fun, will be part of most packages of norms which promote efficient and stable cooperation. Cruelty is no longer offered openly as public entertainment. That is a

change since the days of badger-baiting, and it is a change propelled, in part by the acceptance of an anti-cruelty maxim, and that maxim has been accepted because it is true. The truth of the maxim is not causally idle: it is relevant to its presence, persistence, and learnability.

Even if we accept the view that there are importantly different routes through which moral norms come to be endorsed and internalised, even in favourable cases, there is a striking difference between moral response and reading a bubble chamber photo. When an expert scans bubble chamber photographs, the practitioner knows the key elements of the vindicating history. The trained eye is supported by theoretical reflection. That is not true of intuitive moral response. In their different ways, Hauser and the moral nativists, and Haidt and other sentimentalists have shown that an agent's capacity to vindicate their moral judgement in any coherent terms is often feeble. But even in science, reflective vindication is an achievement of maturity. It is not in place at the beginning of the process. Consider, for example, biological classification before mature evolutionary biology. Linnaeus built on existing practice, but from his work, biological systematics flourished, with organisms being identified, described[9]; sorted in to species, genera, family. By our current lights, these practices were quite reliable. But the practitioners lacked a vindicating theory of their practice; they lacked, for example, a vindicating theory of homologies and how they were to be distinguished from other forms of similarity, though their actual methods were quite reliable. Likewise, an evolutionary theory of the nature of species had to wait until the modern synthesis. The history of systematics shows that it is possible to respond to and track a phenomenon (in this case, the tree of life) without a good account either of the nature of the underlying phenomenon, nor of why the perceptual proxies are in fact good signals of that phenomenon. As we have seen, medical response to malaria was for centuries another case: response to a real phenomenon was not guided by an accurate meta-understanding of that response. It is true and important that those who debated, and continue to debate moral choice often have no good account of offer of the nature of appropriate moral maxims, nor of the evidence that supports one view over another. But the same was true of scientific pioneers. Reflective understanding is an achievement of maturity.

[9] Those descriptions depended, of course, on (on implicit judgments about the traits to be described and those to be ignored; on which traits diagnose differences across kinds of organisms.

5. Is moral knowledge a fuel for success?

Moral language has the form of a fact-stating discourse. "Stalin was cruel"; "Paedophiles deserve to be locked away" have the form of ordinary indicative sentences. That does not show much. Simon Blackburn and Philip Kitcher (for example) have developed theories of moral language and cognition that treat moral language as indicative, but without any serious commitment to moral facts. Kitcher, for example, has defined a notion of moral truth that is parasitic on his core concept, moral progress. More generally, we do not need a robust, correspondence notion of truth to explain the logical or inferential roles of truth: for that, pragmatic or deflationary theories suffice (Horwich 1998). Rather, we need a substantive notion of truth when the representational properties of language and thought help explain success, when that success is flexible; when the representational capacities support adaptive action across a range of projects. An agent's thoughts latch onto something in the world, if having those thoughts is a general asset for the agent's goals in life. We need a robust, explanatory notion of objective fit between mind and world to explain systematic success of thought-guided action; beliefs that accurately represent the world are a fuel for success (Godfrey-Smith 1996; Sterelny 2003). At the end of section 3 we argued that agents who used the framework of ancient astronomy in representing their celestial environment thereby built a mental map that was to some significant degree a fuel for success, despite the theoretical flaws of the framework and despite its incorporation into magical modes of thinking. Are moral beliefs likewise fuels for success, when, and in virtue of, being true?

We see a case for a partially positive answer. But the mixed genealogy of moral thinking is also important. Moral norms often play a dual role of coordinating devices, and as cooperation amplifiers, promoting choices that give other agents incentives to cooperate in turn. These roles can conflict, for once default forms of action establish in a community, agents have incentives to conform to them, even if they eliminate or erode cooperation (Boyd and Richerson 1992). Agents have incentives to match their normative beliefs to those of their community, whether their beliefs are true or not. For adherence to local norms is part of the process that establishes common

knowledge: sets of background expectations about others and how they will behave; expectations on which agents rely in planning and coordination. If local defaults rule out social interaction between the sexes, at best violating those expectations will cause coordination failure and social uncertainty: the agent who does not act as if females were potential sources of pollution is weird, unpredictable. Typically, there are even stronger incentives to conform, because the normative views of an agent are themselves the subject of normative assessment. Moral beliefs are not just reflections of the moral world; they are part of the moral world. Part of being moral is having the right moral beliefs. It is not enough to avoid paedophilia; one must also think that paedophilia is wrong. It is not enough to avoid talking to women; you should think that talking to women is wrong. In contrast, folk astronomy was not especially a tool for coordination and social interaction, and unless they became enmeshed in religion and magic, folk astronomical beliefs were not socially marked. So, there was no special pressure to conform to others' errors.

So moral thinking is not a domain in which, all else equal, the true belief is automatically rewarded. Even so truth — identifying the norms that really do enhance the prospects of profitable and stable cooperation — does power adaptive behaviour in its own right. First: consider the partner choice contexts we considered in section 1: being good to seem good. In contexts of partner choice, social interactions will go better for you, the better you assess the moral facts. You aim to choose, and be chosen by, partners who internalise not just any norms, but rather norms of cooperation, fair-dealing, trustworthiness, commitment to their undertakings. You want such partners even if — perhaps especially if — they are locally unusual. To the extent partnership markets work in ways that defenders of partner choice models suppose, and to the extent that moral commitments are an important aspect of partner value in those markets, the commitments must be of a kind that motivate fair cooperation.

Second, while incentives to conform to any locally dominant norms are real, we should not think of agents as mere passive consumers of the local menu of norms. Agents influence their local normative environments. Most humans now live as globally invisible members of huge societies, but within these vast conglomerates, they live in sets of interconnected microworlds. They live in families, clubs, local

workspaces, informal social groups. Individual attitudes and actions can have significant positive and negative effects on these microworlds. Most of us will have experienced cooperative and friendly microworlds whose character has been formed by the positive influence of a few key individuals. Less happily, most of us have also experienced microworlds whose cooperative dynamics have been ruined. Agents that accept, live and promote prosocial, cooperation-sustaining norms (including the willingness to confront freeloaders) can influence these microworlds in ways that make them better for themselves (and others); better for a wide range of particular plans. True moral beliefs are tools that can help an agent engineer their immediate social environment, even if their global environment is impervious. No doubt the potential to change the local social world beneficially varies greatly from context to context. But we conjecture that it has often been present to some degree.

In our view then, a version of reductive naturalism about moral norms can be built around one perspective on the evolutionary history of moral thinking. Moral truths are principles of action and interaction which support forms of cooperation that are stable because they are fair enough to give almost everyone an incentive to continue to cooperate. In favourable cases, but only favourable cases, these norms are endorsed because they are true, and when endorsed, they support successful social interaction. The vindication is partial. For one thing, moral thinking is not just truth-tracking: it displays community membership and commitment to local mores. To the extent that it is truth tracking, it is error-prone. Our moral views are roughly analogous to the astronomical lore of the ancient world. Just as ancient astronomy was a response to the celestial world, moral views are a response to the opportunities and challenges of a world in which cooperation is profitable but fraught with potentials for conflict and mis-understanding. As in the case of premodern astronomy, these responses do not typically identify and solve those challenges ideally. But in a range of case, individuals and groups normative practices are appropriately shaped by these challenges and the available solutions, and they do enable individuals and groups to act adaptively in their social environments with some reliability. Moral thinking is neither a well-polished mirror of social nature, nor an adaptive fiction.

References

Baumard, N., J.-B. Andre, et al. (2013). "A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice." Behavioral and Brain Science **36**: 59-122.

Boehm, C. (2012). Moral Origins: The Evolution of Virtue, Altruism and Shame. New York, Basic Books.

Bowles, S. and H. Gintis (2011). A Cooperative Species: Human Reciprocity and Its Evolution Princeton, Princeton University Press.

Boyd, R. and P. Richerson (1990). "Group Selection among Alternative Evolutionarily Stable Strategies." Journal of Theoretical Biology **145**: 331-342.

Boyd, R. and P. Richerson (1992). "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." Ethology and Sociobiology **13**: 171-195.

Bulbulia, J. (2004a). "The cognitive and evolutionary psychology of religion." Biology and Philosophy **19**(5): 655-686.

Bulbulia, J. (2004b). "Religious Costs as Adaptations that Signal Altruistic Intention." Evolution and Cognition **10**(1): 19-38.

Chudek, M., W. Zhao, et al. (2013). Culture-Gene Coevolution, Large Scale Coooperation, and the Shaping of Human Social Psychology. Cooperation and Its Evolution. K. Sterelny, R. Joyce, B. Calcott and B. Fraser. Cambridge, MIT Press**:** 425-457.

Churchland, P. (1996). The Neural Representation of the Social World. Minds and Morals. L. May, M. Friedman and A. Clark. Cambridge, MIT Press**:** 91-108.

Dennett, D. C. (1991a). Consciousness Explained. Little, Brown and Company, Boston.

Dennett, D. C. (1991b). "Real Patterns." Journal of Philosophy **87**: 27-51.

Dennett, D. C. (1995). Darwin's Dangerous Idea. New York, Simon and Shuster.

Devitt, M. and K. Sterelny (1999). Language and Reality: An Introduction to Philosophy of Language. Oxford, Blackwell.

Frank, R. (1988). Passion Within Reason: The Strategic Role of the Emotions. New York, WW Norton.

Fraser, B. (2012). "The nature of moral judgments and the extent of the moral domain." Philosophical Explorations **15**(1): 1-16.

Fraser, B. (2013). "Evolutionary debunking arguments and the reliability of moral cognition." Philosophical Studies.

Gauthier, D. (1987). Morals By Agreement. New York, Oxford University Press.

Godfrey-Smith, P. (1996). Complexity and the Function of Mind in Nature. Cambridge, Cambridge University Press.

Godfrey-Smith, P. (2012). Signals, Icons, and Beliefs. Millikan and Her Critics. D. Ryder, J. Kingsbury and K. Williford. Oxford, Blackwell.

Harman, G. (1977). The Nature of Morality. New York, Oxford University Press.

Harris, M. (1985). Good To Eat: Riddles of Food and Culture. New York, Simon and Shuster.

Hauser, M. (2006). Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong. New York, HarperCollins.

Horwich, P. (1998). Truth. New York, Oxford University Press.

Jackson, F. (1997). From Metaphysics to Ethics: A Defense of Conceptual Analysis Oxford, Oxford University Press.
Joyce, R. (2006). Evolution of Morality. Cambridge, Mass, MIT Press.
Lewis, D. (1969). Convention. Oxford, Blackwell.
Mackie, J. (1977). Ethics: Inventing Right and Wrong. London, Penguin Books.
Mameli, M. (2001). "Mindreading, Mindshaping and Evolution." Biology and Philosophy **16**(5): 595-626.
Mikhail, J. (2007). "Universal Moral Grammar: Theory, Evidence, and the Future." Trends in Cognitive Science.
Nichols, S. (2004). Sentimental Rules: On the Natural Foundations of Moral Judgment. New York, Oxford University Press.
Noë, R. (2001). Biological markets: partner choice as the driving force behind the evolution of cooperation. Economics in Nature. Social Dilemmas, Mate Choice and Biological Markets. R. Noë, J. van Hooff and P. Hammerstein. Cambridge, Cambridge University Press**:** 93-118.
Prinz, J. (2007). The Emotional Construction of Morals. . Oxford, Oxford University Press.
Richerson, P. and R. Boyd (2001). "Institutional Evolution in the Holocene: The Rise of Complex Societies." PROCEEDINGS- BRITISH ACADEMY **110**: 197-234.
Richerson, P. and J. Henrich (2012). "Tribal social instincts and the cultural evolution of institutions to solve collective action problems. With Joe Henrich. ." Cliodynamics **3**: 38-80.
Rocco, F. (2000). The Miraculous Fever-Tree: Malaria, Medicine and the Cure that Changed the World. London, Harper Collins.
Ross, D. (2006). "The Economic and Evolutionary Basis of Selves." Cognitive Systems Research **7**: 246-258.
Ruse, M. (1986). Taking Darwin Seriously. Oxford, Blackwell.
Skyrms, B. (2010). Signals: Evolution, Learning and Information. Oxford, Oxford University Press.
Sterelny, K. (2003). Thought in a Hostile World. New York, Blackwell.
Sterelny, K. (2007). "SNAFUS: An Evolutionary Perspective." Biological Theory **2**(2): 317-328.
Sterelny, K. (2010). "Moral Nativism: A Sceptical Response " Mind and Language **25**(3): 279-297.
Sterelny, K. (2012). The Evolved Apprentice Cambridge, MIT Press.
Sterelny, K. (forthcoming). "Symbols, Signals and Norms." Biological Theory.
Stich, S. (1993). Moral Philosophy and Mental Representation. The Origin of Values. M. Hechter, L. Nadel and R. Michod. New York, Aldine de Gruyer**:** 215-228.
Street, S. (2006). "A Darwinian Dilemma for Realist Theories of Value." Philosophical Studies **127**(1).
Wilson, D. S. (2002). Darwin's Cathedral: Evolution, Religion and the Nature of Society. Chicago, University of Chicago Press.
Zawidzki, T. (2013). Mindshaping. Cambridge, MIT Press.