



Darwinian Ethics and Error

R. JOYCE

Department of Philosophy
University of Sheffield
Sheffield
UK
E-mail: r.j.joyce@sheffield.ac.uk

We give to necessity the praise of virtue.

Quintilian, *Institutiono Oratoria*, I, 8, 14

Poor virtue! A mere name thou art, I find,

But I did practise thee as real!

Unknown; cited by Plutarch, *Moralia*, 'De superstitione'

Abstract. Suppose that the human tendency to think of certain actions and omissions as morally required – a notion that surely lies at the heart of moral discourse – is a trait that has been naturally selected for. Many have thought that from this premise we can justify or vindicate moral concepts. I argue that this is mistaken, and defend Michael Ruse's view that the more plausible implication is an error theory – the idea that morality is an illusion foisted upon us by evolution. The naturalistic fallacy is a red herring in this debate, since there is really nothing that counts as a 'fallacy' at all. If morality is an illusion, it appears to follow that we should, upon discovering this, abolish moral discourse on pain of irrationality. I argue that this conclusion is too hasty, and that we may be able usefully to employ a moral discourse, warts and all, without believing in it.

Key words: Darwin, error theory, ethics, evolution, evolutionary ethics, Mackie, naturalistic fallacy, Ruse

Introduction

Michael Ruse argues that morality is fundamentally a product of natural selection, and that the correct metaethical conclusion to draw from this is a moral error theory (Ruse 1986a, 1986b). I am strongly inclined to agree on both counts, and here wish to address some recent opposition. I will not argue for the premise – it will be discussed at the outset only in so far as we need to understand it – rather, it is the movement from the premise to an error theory that interests me here.

Ruse argues that the content of morality is objective – we treat our moral claims as claims about the world. I think that the correct manner of expressing this is to focus on the fact that we consider morality as something that ‘binds’ us, that we cannot opt out of; in other words, the content of morality is that of *categorical* (as opposed to hypothetical) *imperatives*. A hypothetical imperative is the familiar, everyday ‘You ought to catch the 2.30 train’ – the utterer and addressee understand that there is a tacit suffix: ‘... if you want to get to so-and-so in good time’. If it turns out that the addressee lacks that end, then the imperative is withdrawn. A categorical imperative, by contrast, ‘declares an action to be objectively necessary in itself without any reference to any purpose’ (Kant 1993: 78). Categorical imperatives can be seen as ‘about the world’ inasmuch as they apparently appeal to rules of conduct ‘which are simply there, in the nature of things, without being the requirements of any person or body of persons’ (Mackie 1977: 59).¹ It seems quite correct that moral discourse is objective in this sense: when we condemn a moral villain, we do not first check that he has the appropriate interests or desires. If he is guilty of something repugnant, like stealing from innocent people on a whim, we would not dream of retracting our judgment ‘He ought not do it’ upon discovering that he has a conflict-free desire to steal (and desires all likely consequences of stealing too); there is nothing he can assert (however truly) concerning his ends and interests that will get him off the hook.

From an evolutionary point of view, there is a good explanation for our treating morality as consisting of categorical imperatives. The actions that morality prescribes with categorical force are those that constitute or promote, roughly speaking, *cooperation*. To cooperate with those who may return the favour (reciprocal altruism), and those who share a substantial portion of one’s genetic material (kin altruism), enhances reproductive fitness. Therefore evolutionary forces have favoured cooperation.² Evolution might have simply ‘made us’ cooperate (and refrain from defecting), or might have granted us powerful epistemic abilities whereby we can calculate the reproductive advantages of cooperation on a case by case basis. But neither option is optimally efficient. Of the former, Ruse writes: ‘we would have wasted the virtues of our brain power, and the flexibility which it gives us’; of the latter: ‘this would have required massive brain power to calculate probabilities and the like’ (Ruse 1986a: 221). A ‘middle road’ is selected for: we have evolved an innate disposition in favour of certain types of action, against certain others. This disposition is not merely the development of appropriate *emotions* or *desires*: it’s not merely that I *want* to look after my children – but I feel that I *ought* to. I feel, if you will, that there is a *requirement* upon me to look after my children; that I *must*. Desires, after all, are unreliable things: after a long day, a parent might not particularly *want* to care for the children,

and this is where a sense of *requirement* kicks in. Since the desire is absent – since the long-term satisfactions of child-rearing are being under-appreciated due to distraction, weakness of will, or simple exhaustion – it is important that the requirement is not conceived of in hypothetical terms. Morality as a system of categorical imperatives compensates for the limitations of desire.

Though the above raises a great many questions and, no doubt, objections, here I wish neither to defend it nor elaborate it, but rather ask, from a metaethical point of view, what follows from it. My contention is that Ruse is correct in holding that the most plausible consequence is a moral error theory. There do not really exist *any* categorical requirements binding our actions, enjoining cooperation and proscribing defection. It's all an illusion which, in evolutionary terms, has served us very effectively. Thus all our judgments of the form ' ϕ is morally obligatory' are untrue: *moral obligatoriness* is a property that no actual action instantiates.³ In practical terms, cooperation is fostered most effectively if we have a disposition to see it as categorically required: 'morality simply does not work . . . unless we believe that it is objective' (Ruse 1986a: 253). But, in metaphysical terms, there is no need to think that there *are* such requirements: everything that needs explaining is explained by the thesis of evolutionary error. The further hypothesis, that these judgments are *true* – that there is a realm of moral facts – is redundant. (And if Ockham's Razor doesn't do the trick, then categorical imperatives can be tackled head-on: Mackie argues that they are 'queer'; Philippa Foot argues that they depend for their legitimacy on 'a magical force', etc. (Foot 1972))

As a final preliminary, let me stipulate a distinction between a moral error theory and what I shall dub 'moral abolitionism'. A moral error theory states that our moral judgments are fundamentally flawed – that our moral discourse contains few, if any, true judgements. But an error theory does not entail anything concerning what we ought to *do* with our discourse once we've uncovered its flaws. (Note that the previous 'ought' did not pose as a *moral* 'ought', so there is no circularity lurking in the question 'Given that there is nothing that we morally ought to do, what ought we to do?') Moral abolitionism is one way of answering that question: it is the view that we ought to *do away* with it. Thus the error theoretic stance is a philosophical position, whereas what I am calling 'abolitionism' is the result of a practical decision. That one leads to the other is a natural enough thought. Elizabeth Anscombe – believing that our moral deontological concepts (concerning what we *ought* to do, what we *must not* do, etc.) are 'survivals, or derivatives from survivals, from an earlier conception of ethics which no longer generally survives', and are unintelligible outside that framework – concludes that they must 'be

jettisoned if this is psychologically possible' (Anscombe 1958). But the move from error to abolition is by no means mandatory, and Ruse stoutly resists it:

I hasten to add that I am not now suggesting that morality is in any way a sign of immaturity. Nor would I have those of us who see the illusory nature of morality's objectivity throw over moral thought. . . . Morality is a part of human nature, and . . . an effective adaptation. Why should we forego morality any more than we should put out our eyes? I would not say that we could not escape morality – presumably we could get into wholesale, anti-morality, genetic engineering – but I strongly suspect that a simple attempt to ignore it will fail. This is surely the (true) message of Dostoevsky in *Crime and Punishment*. Raskolnikov tries to go beyond conventional right and wrong, but finds ultimately that this is impossible (1986a: 253).

I will discuss the move from error to abolition in the final section, but first I shall address the question of the passage from morality being an evolved trait to morality being in error.

Evolutionary ethics and success

Earlier enthusiasts of 'evolutionary ethics' sought in natural selection a *vindication* of a kind of morality: *moral goodness* might be identified with (something like) *is [or has been] naturally selected for*. Since there is a plausible case to be made that certain types of action and psychological trait have been naturally selected for, there is a plausible case to be made that certain actions and traits are morally good. The error theory disappears, to be replaced with an evolutionary *success* theory! Ruse will have none of this, and is particularly sensitive to the concern that any such theory will fall foul of the *naturalistic fallacy* (of which, more later). Nevertheless, several commentators, accepting that some of the attitudes we have towards cooperative actions are born of natural selection, still think that a kind of evolutionary success theory is on the cards. (Just to be clear, by 'a success theory', I mean one that holds that our moral discourse is not fundamentally in error, that many of our utterances – such as ' ϕ is morally wrong', 'You must ψ ', etc. – are true. An *evolutionary* success theory shall hold that the kind of fact in virtue of which such judgments are true is, in some manner, a fact about human evolution.)

Ruse, in Humean spirit, sees morality as a matter of our 'objectifying' our moral sentiments (Ruse 1986a: 253). Says Hume: 'Vice and virtue may be compared to sounds, colours, heat and cold, which, according to modern philosophy, are not qualities in objects but perceptions in the mind' (Hume

1978: 469) – moral judgments are a matter of the ‘gilding and staining [of] natural objects with the colours borrowed from internal sentiment.’ (Hume 1983, Appendix I). Regarding the ontology of colour, it has been a popular strategy in recent years to accept Hume’s basic projectivist premise, yet to place colours in the world, as a dispositional property of the surfaces of objects. Redness, for example, is said to be the dispositional property of producing the phenomenological response *redness* in normal human viewers (as they are actually constituted) under good viewing conditions (i.e., in broad daylight) (McDowell 1985; Johnston 1992; Campbell 1993). There is a kind of objectivity here, since had a tomato ripened fifty million years ago it would still be red, in so far as *were* a normal human to observe it in good viewing conditions (never mind that there weren’t any humans in existence) that human *would* have a certain response. We might say that this analysis makes colours existentially independent of, though conceptually dependent on, human minds.

William Rottschaefer and David Martinsen attempt the same move for morality: we can accept that positive attitudes towards cooperative actions have been naturally selected for, yet identify *moral rightness* (for example) with a relational property instantiated by these actions: (something like) *such that humans have evolved to respond with favour* (or even: *such that humans have evolved to have a response of ‘moral objectification’*) (Rottschaefer and Martinsen 1990; Rottschaefer 1998). Now if we’re accepting the premise that the attitude favouring cooperative activity is an evolved trait, then it cannot be denied that such activity *does* instantiate the kind of relational property gestured at, but a crucial question remains: ‘Is that property the referent of the term *rightness*?’ Regarding colour, the point is put succinctly by Michael Smith (1993: 239): ‘Someone who denies that colours are properties of objects need not deny that objects *have* these dispositions, all he has to deny is that colours *are* such dispositions.’ The mere availability of a dispositional account of a concept does not force that analysis upon us. After all, for *any* predicate we can find a dispositional property had by all and only the items in the predicate’s extension. All and only the objects satisfying ‘... is a manatee’ are (trivially) such that they *would prompt the response* ‘*There’s a manatee!*’ in an infallible manatee spotter.

Let’s allow that cooperative actions of a certain kind have a ‘Darwinian’ dispositional property – they are such that humans have, through the pressures of natural selection, come to favour them. (That may be vague, but it’s adequate for our general purposes.) Would there be reason to *resist* thinking that this property is the referent of a familiar moral term of positive appraisal? Yes. For such a property cannot (at least as far as I can see) underwrite the notion of moral *requirement* – and what is moral rightness, if not something

we are *required* to pursue? Consider again the unrepentant moral villain earlier mentioned. We can allow that the action he performed had the following relational property: being such that humans have evolved to respond with disfavour. According to Rottschaefer and Martinsen, then, the action was *wrong* – really, objectively wrong. Unfortunately, moral naturalism does not come that easily. For at the heart of our moral discourse is the idea that the criminal *ought not* to have performed the action, that he was somehow *required* to refrain. And it would be very odd if we thought that he ought to ϕ while admitting that he has no reason to ϕ ; therefore canons of ordinary moral thinking will also suppose the criminal to have had a *reason* to refrain (regardless of his desires, and regardless of whether he is aware of the fact). But why do the things favoured by natural selection *bind* him, or provide him with *reasons*? Moreover, many have thought that if a person makes a moral judgment (that some action is wrong), it follows of necessity that she has some *prima facie motivation* against that action.⁴ But the criminal may note with utter indifference that an action is such that humans have been naturally selected to disfavour it – what’s that fact to *him*?

It would be tempting, but futile, to appeal to the fact that our criminal *is* a human, with all the natural human dispositions, and therefore has reason to act in accordance with natural selection. This is, in effect, how Robert Richards argues in presenting his evolutionary success theory (Richards 1986). Since, according to Richards, all humans have evolved to act for the community good, we may say to any human: ‘Since you are a moral being, constituted so by evolution, you ought to act for the community good’. He likens this derivation of an ‘ought’ to that occurring in ‘Since lightning has struck, thunder ought to follow’. This is surprising, since the ‘ought’ of the latter is an *epistemic* or *predictive* one. Such ‘ought’s still, arguably, entail reasons: ‘That lightning has struck gives one reason for believing that thunder will follow’ (Harman 1975; Mackie 1977: 74). But the *moral* ‘ought’ that Richards hopes to derive surely is not an epistemic one: when we say that the villain ought not steal, we are not saying that we are able to predict, on the basis of some antecedent concerning evolution, that he will not steal; and, by the same token, the *reason* entailed by the ‘ought’ pertains to *his* reasons for not stealing, not *our* reasons for believing that he won’t steal!

Presumably what Richards hopes to do is to make moral imperatives *hypothetical*, depending for their legitimacy on an end which all humans, as a matter of fact, have been assigned by natural selection: the good of the community. If our moral villain has this end, then he ought to do (*ceteris paribus*) whatever will satisfy it; he has a (*prima facie*) reason to do whatever will satisfy it. Now evolutionary forces have certainly not bestowed upon us all an active *desire* to promote community good – at most, we are endowed with a

disposition, or *capacity*, in favour of its promotion (as Richards recognizes). But why does a mere disposition provide an ‘end’, or ground an ‘ought’ statement? In general, ‘oughts’ may be grounded by desires – if Sally wants coffee, then, *ceteris paribus*, she ought to head to the café – and *interests* (if they are distinct from desires) may also ground ‘oughts’ – if it is in Sally’s interests to stop smoking, despite her having conflict-free desires to carry on, then, *ceteris paribus*, she ought to stop. But I cannot see that the same goes for dispositions. Allow that evolution has endowed Jack with a disposition to favour the promotion of the community’s good, but imagine that his upbringing was such that the disposition went quite undeveloped, and now has been effectively quashed. Why *ought* he still act for the community’s good? Why does he still have a *reason* to?⁵

Richards toys with the idea of simply branding Jack a ‘sociopath’, therefore not fully human, and therefore not a proper subject of moral injunctions. Perhaps this would stick if Jack lacked the disposition altogether, as the result of a genetic aberration, but we are not claiming any genetic anomaly – Jack still *has* the disposition, it has just gone utterly undeveloped, and now, let’s imagine, it is too late for Jack to develop it, in much the same way as it is now too late for him to become a concert pianist.⁶ It’s important to note that our ‘villain’, despite earlier characterizations, need not be the serial killer stalking back streets, need not be the suicidal teenager heading to school with an automatic rifle in his bag. The kind of self-centred person we encounter every day – one who regulates his or her actions consciously and solely in terms of perceived self-gain – will suffice perfectly well as an example of someone whose altruistic dispositions have been quelled. Bearing this in mind, talk of ‘sociopaths’ who fall short of satisfying the criteria for *being human* seems wildly overstated.

Consider such a character: pleasant enough to interact with, has a successful career, a family, etc. But if she has made a promise that will be inconvenient to keep, and she sees that she can break it without incurring penalty (perhaps she can make a decent excuse), then, despite her knowledge that doing so will seriously penalize others, and, say, harm the community in general, she will not hesitate to break the promise. Let us point out to her that the action of promise-breaking has a certain ‘Darwinian’ dispositional property: it is such that humans have evolved to disfavour it. She accepts this, but notes it with unconcern (along with facts about the evolution of manatees). Let us inform her that she herself has this disposition, in the sense that had she received a certain kind of upbringing she would have favoured the good of the community (and may pass this disposition on to her offspring). But, given that she *didn’t* receive that upbringing, but one that left the disposition dormant, why does she now have a reason to refrain from promise-breaking? To say

that the disposition *must* have some manifestation, such that in some sense she, in acting against the community's good, *must* be subtly undermining her own projects and interests, is just desperate.

These observations disclose my doubt concerning Ruse's view that morality is unavoidably with us, and that only genetic engineering could eliminate it (see his quote about Raskolnikov above). If the employment of moral concepts is a genetically present *disposition*, then it is perfectly possible that certain socialization processes (or perhaps merely a course of metaethics) could dampen or completely nullify the moral sentiment. In my experience, there are more people around who do not properly participate in moral thinking (but who are hardly thereby 'psychopathic') than philosophers like to admit. If this is correct, then it would be possible to eliminate moral discourse without resorting to genetic tampering if we wanted to (as we can, arguably, disable the manifestation of aggressive or xenophobic dispositions). Whether we *ought* to do so is the subject of Section IV.

I have re-iterated the question of why facts about evolution provide persons with *reasons*, why they ground moral 'ought' statements. And it should be clear that my answer is: 'As far as I can see, they don't.' Of course, if evolution has endowed me with a disposition to favour cooperation, and my upbringing was such that this disposition *has* developed fully, then indeed I have a (*prima facie*) reason to cooperate. But now all the work is being done by the fact that my upbringing provided me with certain attitudes and traits that are now actively operative – and these attitudes would ground 'ought' statements even they had nothing to do with evolution. It will not do to maintain that any agent in whom such dispositions lie untapped (and now 'untappable') is simply a *sociopath*, who lies beyond the pale of moral injunctions. We have already seen that such agents are possibly quite common, and they certainly remain subject to the dictates of moral discourse. We think that a person – regardless of an upbringing that left her intractably selfish – morally *ought not* break promises for the sake of convenience. Pointing to a relational property pertaining to natural 'fitness', indicating that natural selection provides humans with certain dispositions against promise-breaking, does not help. And if an ethical theory cannot account for so central and familiar a moral judgment – that a selfish person ought not break an inconvenient promise – it has not gotten off the ground.

Robust Darwinian naturalism and the naturalistic fallacy

Rottschaefer and Martinsen anticipate an accusation from Ruse that their theory blunders into the dreaded naturalistic fallacy, and go to some effort (as does Richards) to show that it does not. But it is not the naturalistic fallacy

that I accuse such theorists of, for, I hereby admit, I have little idea what that fallacy is, nor why ethicists – especially those interested in evolution – seem so fearfully mesmerized by it.⁷ It has become commonplace to assume that G.E. Moore's notorious fallacy does for 'good' what Hume did for 'ought', but no part of *Principia Ethica* that I am familiar with bears resemblance to Hume's claim that one cannot derive an 'ought' from an 'is' (Moore 1903). It is true that nothing like the following is formally valid (if by this we mean 'is an instance of a theorem of the predicate calculus'):

- (1) Things of type ϕ are such that humans, by the process of natural selection, are disposed to have attitude A towards ϕ .

Therefore: Things of type ϕ are morally good.

But no naturalist would claim such a thing. Rather, she will treat the above as an enthymeme, inserting a major premise if required:

- (2) If things of type ϕ are such that humans, by the process of natural selection, are disposed to have attitude A towards ϕ , then things of type ϕ are morally good.

It is not good complaining that (2) reproduces, in conditional form, a formally invalid argument, for the naturalist does not claim that (2) is 'valid', merely that it is *true*. Nor can it be simply insisted that (2) commits 'the naturalistic fallacy' in virtue of relating a fact to a value, and therefore must be false. That's just begging the question. It is also important to remember that the 'fallacy', according to Moore, is committed no less by statements of the following kind:

- (Yellow) Having the natural properties P, Q, R, *is* what it is to be yellow.

So he evidently did not think that it is the 'evaluateness' of goodness that powers the fallacy, but its *indefinability*. But again, we cannot simply *assume* that goodness is indefinable (or unanalyzable), for that is precisely a point at issue. When we look at the heart of Moore's description of the fallacy (in §12), what we actually find seems to be advice that we ought not confuse the 'is' of identity with the 'is' of predication. Moore thinks that the hedonic naturalist, when he claims 'Pleasure is good' may be saying something true so long as it's an 'is' of predication; but to mistake it for an 'is' of identity (a *definition*, by Moore's lights) leads to absurdity. In the same way, if I say 'The book is red' and 'The book is square' – but these are taken as identity claims – I'm left with the crazy conclusion that redness is squareness.

Keeping track of one's 'is's is surely good advice – perhaps to confuse them may even be called a kind of 'fallacy' – but Moore is quite mistaken if he thinks that the naturalist *must* be confused over 'is'. (2) can be seen as entailed by a naturalistic thesis:

(Naturalism) For any ϕ , ϕ is a type of thing towards which humans, by the process of natural selection, are disposed to have attitude A if and only if things of type ϕ are morally good.

There is one 'is' of predication there. With rewording, the biconditional might be strengthened into an 'is' of identity flanked by property names. Thus naturalism might be an *a posteriori* claim, comparable to 'Water is H₂O', or an *a priori* (but covert) thesis, like 'Knowledge is justified true belief'. But in neither case need the naturalist fall foul of the problem that Moore called 'the naturalistic fallacy'. Moore does allow that *some* things may be defined without trouble: his stock example is a definition of horse. (That's Moore's syntax; I'd much rather speak of a definition of 'horse' or a definition of *horseness*. Since he's adamant he does not intend the former, I assume he means the latter.) So what is it about *yellowness* and *goodness* that makes them different from *horseness*? Moore's answer is that they are 'simple', 'non-natural', and 'indefinable' – but this cannot be treated as a self-evident datum, for it is exactly what the naturalist, in offering something like (Naturalism), denies. Oddly, of this all-important premise, Moore writes: 'As for the reasons why good [sic] is not to be considered a natural object, they may be reserved for discussion in another place.' It appears that this 'further discussion' is the very next section of *Principia Ethica* – where the Open Question Argument is deployed. But the woes of that argument are well-documented, and won't be rehearsed here. For effective criticism, see, e.g. Harman (1977: 19), Frankena (1973: 99ff.) and Putnam (1981: 205ff.). (I'll merely note that it doesn't even work for Moore's favourite example of the definition of *horseness* – for the analysis he offers is *a posteriori* in nature – making mention of a horse's *heart* and *liver*, etc. – such that a perfectly competent speaker might be certain that X is a horse, but uncertain that X has property N [where 'N' stands for the 'definition' Moore offers, involving hearts and livers].)

Consider something like (Naturalism) – what Rottschaefer and Martinsen would call a 'robust Darwinian naturalism' (and I have called 'an evolutionary success theory'). The question, I have argued, is not whether it commits a 'fallacy', but whether it is *true*. If it is true, then it is either an *a priori* or an *a posteriori* truth. The relevant model for the former is a philosophical analysis like 'Knowledge is true justified belief'. We do not

come upon such truths (pretending that it *is* a truth) simply by doing a bit of quick introspection, or by looking in a dictionary. Smith suggests that one way of proceeding is to gather all our platitudes about knowledge – a platitude being something one comes to treat *as* platitudinous in attaining basic competency with the concept – and then to systematize those platitudes (Smith 1993, 1994). ‘True, justified belief’ may be the best systematization, or encapsulation, of our epistemic platitudes (though it probably isn’t). But it is clear that no description worded centrally in *evolutionary* terms is going to be the best systematization of our moral platitudes. Moral concepts, I assume, preserved their identity criteria throughout the nineteenth century: someone saying ‘Slavery is morally wrong’ in 1890 was not expressing a different proposition to someone uttering the same sentence in 1810 (otherwise, were the 1810 speaker instead to assert ‘Slavery is morally permissible’, she would not be in disagreement with the 1890 speaker, in which case we could not say that moral attitudes towards slavery changed over the course of the nineteenth century). If this is true, then, according to the theory under question, it was *a priori* available to pre-Darwinian speakers to systematize their moral platitudes in such a way that *natural selection* centrally figured in that explication. But that is absurd, so robust Darwinian naturalism as an *a priori* thesis is a non-starter.

How will it fare as an *a posteriori* thesis? The model here is ‘Water is H₂O’. According to the *a posteriori* naturalist, we can ‘find out’ that two kind terms, perhaps both in common parlance, are, and always have been, co-referential. See, e.g., Boyd (1988) and Brink (1984). This sounds closer to what the robust Darwinian naturalist will presumably claim: when we consider a term like ‘moral rightness’, and examine the kind of things to which we apply it (and the kind of things from which we withhold it), and then bring in evolutionary theory, perhaps boosted by detailed empirical confirmation, we might discover that (pretty much) all and only the things to which we apply ‘...is right’ instantiate a property, or cluster of properties, which may also be described by the predicate ‘...is a type of thing towards which humans, by the process of natural selection, are disposed to have attitude A’. This is potentially threatening to an evolutionary error theory, for we have agreed that there *is* such a property had (pretty much) by all and only the things to which we apply our predicate ‘...is morally right’, so is that not immediately to give the game to the (*a posteriori*) evolutionary *success* theorist?

I think not. The worry with this kind of *a posteriori* theory is that it threatens to achieve far too much. Consider our term ‘witch’, that was once applied to actual persons. It is possible that all and only the persons to whom we applied ‘witch’ had a certain property, or cluster of properties – perhaps

they were women who tended to be of a certain social class, playing a certain socio-political role, who threatened the patriarchal authorities in a particular way (I'm not suggesting that it's anything so simple – it may be disjunctive and vague). But to locate such a property clearly would not be an *a posteriori* vindication of 'witch discourse'. Similarly, we have a term 'phlogiston' that we used to apply to various phenomena: we could point to any open flame and say 'Look, there's the phlogiston escaping.' In recognizing that there clearly is a property, or cluster of properties, that all and only open flames have, have we thereby rescued phlogiston discourse? The reason that the answer is obviously 'No' is that when seventeenth century speakers used the predicate '...is phlogiston' (or '...is a witch') something more was going on than merely applying it to some objects, withholding it from others. What doomed the predicate to emptiness, despite its ostensive paradigms, was that users of the term (considered collectively) thought and said certain things *about* phlogiston, such as 'It is that stuff stored in bodies', 'It is that stuff that is released during combustion', and these concomitant statements are false. (The analogous claim for witches will concern their supernatural abilities.)

In my opinion the same thing will go for moral discourse. It is not enough to find some property had by all and only the things to which we apply our moral terms. There are also very important things which we endorse *about*, say, morally right actions – such as they are the ones which a person *ought to* perform regardless of his desires, they are the ones that we have overriding *reason* to perform, they are the ones the recognition of which will *motivate* an agent. But, as I argued previously, the kind of property adverted to by the robust Darwinian naturalist does not satisfy such a sense of 'inescapable requirement' (or, at least, it will require a great deal more argument to show that it does – the prospects for which I am very skeptical of). Therefore this Darwinian dispositional property, though very probably existing, does not deserve the name 'moral rightness'.

The naturalist might respond: 'So much the worse for our sense of *categorical imperative*. Why not just admit that this aspect of our moral discourse is faulty, and carry on with a revised naturalist discourse?' Well, when Lavoisier gave us oxygen theory in the late eighteenth century, why couldn't the fans of phlogiston just revise their theory, insisting that they had been talking about *oxygen* all along, concerning which they had held some false beliefs about its being stored and release? (Ditto, *mutatis mutandis*, for witch discourse.) The reason that it was *not* available for them to revise and vindicate phlogiston theory in this manner is that the thesis about phlogiston being stored in bodies and released during combustion was too central to the theory to be negotiable – one might say that the whole point of phlogiston discourse was to refer to a *stored and released* material. By the same token, I believe, the whole

point of having a moral discourse is to prescribe and condemn various actions with *categorical* force. We have a moral discourse so that various actions (and omissions) can be demanded when desires (whether self-interested or otherwise) are absent, limited, or fail to motivate. If this were not the case, why did we develop a moral discourse at all? – after all, we’ve always had a perfectly well-structured vocabulary for discussing the means of satisfying desires and fulfilling ends – even long-term ones. Evidently, the language of hypothetical imperatives was not adequate to the task for which we required moral language.

Let me sum up this section before moving on to a rather different topic. The robust Darwinian naturalist – he who agrees that various moral attitudes are the result of natural selection, but hopes to found upon this a moral *success* theory, a kind of moral realism – fails to accommodate some very central moral beliefs. I think there are several fundamental desiderata that will go unsatisfied, but here I have focused on the notion of a categorical imperative. In doing so, the naturalist does not commit any form of ‘fallacy’ – he merely presents a false theory. Clearly, there are two vital premises to my position that have only been sketched in a rather brief and dogmatic manner: (i) that our moral discourse *is* centrally committed to categorical imperatives, and (ii) that the robust Darwinian naturalist cannot accommodate these imperatives in his system. Successfully combating either claim would undermine much of what I have said.

One might well wonder what work is now being done by the thesis that our sense of categorical requirement is a *biological adaptation*, for if we can show that moral discourse is centrally committed to thesis T, and that T is philosophically indefensible, then we have our error theory right there, with no mention of evolution. This is, of course, exactly what Mackie and others have tried to do – to establish a moral error theory head-on. But any error theorist owes us an account of *why* we have all been led to such a drastic mistake; the absence of such an explanation is likely to raise doubts that we *are* making a mistake at all (i.e., either doubts that T is erroneous, or doubts that our discourse ever committed itself to T). This, I believe, is where an evolutionary account of the development of moral sentiments plays its role. In other words, we have two theses: one is the error theoretic stance for moral discourse, the other is the claim that morality is largely the product of natural selection. The former, by itself, lacks persuasiveness – it lacks an explanation of where the error came from. The claim that morality is an evolutionary trait – that developing a sense of ‘intrinsic requirement’ would be beneficial to humans *even if there were no such thing* – fills that gap. But the latter thesis, by itself, is insufficient to establish an error theory. I mentioned earlier the possibility of arguing for an error theory using Ockham’s Razor: everything

that needs explaining is explained by an evolutionary story concerning how and why we have a disposition to make moral judgments, with no need for an additional theory according to which the judgments are *true*. But it can now be seen that Ockham's Razor won't suffice, for the kind of robust Darwinian naturalism that has been under discussion does not posit any extra *ontology* – it rather points to dispositional properties, the existence of which all parties to the debate should antecedently agree to.⁸ So Ruse can plausibly claim that we have evolved to believe in objective requirements, but no investigation of the processes of natural selection, of the course of human evolution – no matter how subtle and empirically well-confirmed – will be sufficient to establish that we are victims of *an illusion*. For that we need philosophical argumentation.

Error, abolition and acceptance

Thus far I have argued that those who hope to find in the (probable) fact that certain attitudes have been naturally selected for a *vindication* and *justification* of moral discourse are backing a worthy but misguided cause. Moral injunctions have an *authority* that evolutionary facts cannot underwrite, and being able to appeal to such an authority is the whole point of having a moral discourse. But this authority may yet be shown to be justified in some other manner – certainly there are well-developed philosophical programmes that seek to substantiate it. My judgment is that none of them will be fruitful. Here is not the place to defend that skepticism, but in the remainder I want to investigate what would follow if we decided that the skepticism *is* well-founded – if it is true that we have evolved to accept an illusion, as Ruse thinks.

Let me open discussion with two quotes from earlier articles in this journal. William Hughes writes: 'if [moral values] are unreal then the only rational position is to seek to eradicate moral and ethical language altogether, and replace it with the language of needs and wants (Hughes 1986: 306). And Peter Woolcock, in a cogent critique of Ruse: 'Once we realise [that there are no moral obligations whatsoever], the rational course would seem to be to train ourselves out of any residual tendencies to obey moral laws where we can get away with breaking them. We should deprogramme ourselves out of any inclination to feel guilt, or to want redemption. Contrary to Ruse's denial Nietzsche and Thrasymachus were right – moral thought is overthrown (Woolcock 1993: 428).

A consequence of Ruse's view is that no statement of the form 'S is under a moral obligation to ϕ ' is true. Thus no belief having that content is true. Thus, if a person has evidence of this fact – once 'the cat is out of the bag'

(as Woolcock puts it) – to have such a belief is irrational. Thus, if we read Ruse's 1986 book and justifiably believe it (or, for that matter, if we read Mackie's relevant work and justifiably believe it), it becomes *irrational* for us to hold moral beliefs. I do not see that Ruse can avoid this conclusion without revising his basic position. It might seem that Hughes' and Woolcock's 'abolitionism' follows close on the heels of this admission, but this is exactly what I want to resist. Moral discourse may still have an active role to play even for those who have seen the cat out of the bag.

One way to proceed (that I don't favour) would be to argue, seemingly paradoxically, that it may sometimes be rational to be irrational. There are different things that admit of the 'rational/irrational' distinction – actions, beliefs, emotions – and it is far from obvious that all are appraised for rationality according to the same framework. For example, certain phobic emotions are deemed irrational, often on the grounds that they are experienced in the presence of inappropriate beliefs (I know the spider is harmless, but it fills me, nevertheless, with dread). Suppose, however, that a person is in an unusual situation, such that having a phobia is greatly to her advantage (perhaps she is developing a worthwhile and loving relationship with her therapist). Suppose, moreover, that there is some action that she can perform that will encourage the development of that phobia (she goes to see the movie *Arachnophobia*, knowing it will traumatize her). Since that action is to her instrumental advantage, and she knows it, we ought to deem it rational; the phobic emotion of fear, however, remains no less irrational. So is *she* rational or irrational? The correct answer, it seems to me, is that according to one normative framework she is rational, according to another she is not. The 'post-Ruseian' moralist may be in the same situation. His ongoing belief in moral obligation is irrational, yet his having that belief may be to his practical advantage, may serve his ends, and therefore if there are actions he can perform to encourage such beliefs, those actions are rational.⁹

As I say, I do not favour this kind of defence, encouraging, as it apparently does, a kind of schizophrenia, or self-deception, in the agent. Besides, Ruse has not shown that it is to any *individual's* instrumental advantage to have moral beliefs, only that having moral beliefs has enhanced reproductive fitness. A group of humans who find in cooperative actions a 'to-be-done-ness' does relatively well, their society flourishes, and their genes are passed on. But of any individual living in such a society we can see that *his* advantage is to defect on promises when he can get away with it. The fact that things would go badly for him if *everyone* thought this way is nothing to him – it merely means that he has to try to encourage moral beliefs in others. Nor does the fact that the trait of 'being a free-rider' is unlikely to be favoured by

natural selection in social creatures alter its being to *his* advantage to get a free ride if he can.

Hume (as in so many things) has some interesting thoughts on free-riders (1983, §IX, part 2). First, he points out that there are important values that the free-rider misses – values that by their very nature cannot be gained through secret defection: the satisfaction of fair dealing, comradeship, open cultural participation, etc. Second, free-riders are certainly epistemically fallible, and possibly weak of will, thus ‘while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give into the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all trust and confidence with mankind’. Those looking to defect secretly are likely to miscalculate, get caught, and pay a serious price; if they are also weak of will then the likelihood increases.

It might be argued that what follows from Hume’s observation is that clear-headed calculations of expected self-gain will suffice to regulate cooperative behaviour, with no troublesome *moral duties* or *categorical imperatives* entering into the picture at all. The ‘sensible knave’ would break a promise if she could be sure of getting away with it, but she is sensible enough to know that she is rarely sure of getting away with it, and the price of detection is too great to risk it. Therefore, in all but unusual cases, enlightened self-interest will serve to underwrite all the prescriptions that we would usually call ‘moral’. Moral injunctions may be replaced, after all, by ‘the language of needs and wants.’

But I don’t think that this would be the correct moral to draw from Hume. If Hume’s ‘knave’ really is sensible, he knows that he is epistemically fallible and vulnerable to weakness of will. He knows that the profits of short-term gain are often tempting. He knows, furthermore, that, being human, he is a creature of habit, so a single successful defection might encourage other riskier defections. It is therefore to his advantage to regulate his day-to-day decision procedures by something other than clear-headed egoism, if only because egoistic calculations – as anyone knows who has ever taken up an exercise programme, or embarked on a diet – do not guarantee correct behaviour. What this knave needs is to place a strong value on certain actions, and a strong disvalue on others. He needs to think of cooperative behaviour not in terms of ‘This will, in the long-term, be to my benefit – I just shouldn’t risk defecting; someone might be watching’; rather, he needs to think of it as ‘*This must be done*’. When he takes this step, then he is a knave no more. Of course, employing such a moral concept will not *guarantee* correct behaviour either, but it stands a much better chance.

It might seem that we have argued in a circle: we are back to claiming that having moral beliefs is to the advantage of a standardly situated agent, but

we have not dispelled the fact that to believe p while having been exposed to evidence that firmly discredits p is to be irrational. My way out of this circle (and I offer it as a defense of Ruse) is to deny that getting the regulative benefit from moral concepts requires their figuring in *beliefs*. Think of how we best fend off akrasia when commencing a programme of exercise. I tell myself that I *must* run for an hour every other day (that's just a round number; I don't pretend to achieve anything so impressive!). Of course, it's false that I must run this much and no less: if occasionally I run for fifty-five minutes, or occasionally skip a few days, I'll still achieve my goal of fitness perfectly adequately. But the spirit is weak! – if I start allowing these little lapses, the slippery slope of self-sabotage beckons. What keeps me on track for my goal is a firm and non-negotiable rule: an hour every other day, no less. However, I do not need to *believe* this rule for it to work – if someone questions me, suggesting that there's no harm in occasionally skipping a few days, I am not committed to arguing that this is mistaken – what's important is that I rehearse the rule in my mind, that I allow it to influence my actions, that I let it carry weight with me. I *accept* the rule, but I do not believe it. Indeed, if you were to press me seriously about its truth – in a critical context, *not* when I am actually running – then I would happily express my *disbelief* in it.¹⁰

There is more that we can do with a false theory than either irrationally believe it or abolish it entirely. As a useful fiction it can still have a practical role in our lives (as, indeed, literary fictions have a practical role in our lives). This, I believe, is an option that is available to us concerning *morality* even after we realize that its central concepts are illusions foisted upon us by natural selection. It remains practically advantageous for any ordinarily situated individual to imbue certain cooperative actions with a sense of 'inescapable to-be-done-ness'. It is *more* advantageous for her to do this (I am suggesting) than merely to believe that the same action ought to be performed because it is in her long-term best interest (though she may well believe this as well); and for her really to *believe* that those actions 'must be done' – after reading and justifiably believing Ruse and Mackie, that is – that too would be practically disadvantageous: to believe things the evidence of whose falsehood is available to us is irrational, and is likely to have serious detrimental consequences if adopted as a doxastic policy.

Wittgenstein (1965) once remarked that our moral discourse seems to consist largely of similes. I am reminded of Bentham's slightly bizarre attempt to analyze the idea behind *obligation*: 'the emblematical, or archetypical image, is that of a man lying down with a heavy body pressing upon him (Bentham 1843: 247), as well as Mackie's talk of obligation being an 'invisible cord' and a demand for payment being an 'immaterial suction-pipe' dredging for

the owed money (Mackie 1977: 74). ‘But,’ Wittgenstein continues, ‘a simile must be a simile for *something*. . . [Yet] as soon as we try to drop the simile and simply state the facts which stand behind it, we find there are no such facts. And so, what at first appeared to be a simile now seems to be mere nonsense.’ Though Wittgenstein concludes that the ‘very essence’ of morality is its nonsensicality, he does not advocate its abandonment: it is something he ‘cannot help respecting deeply’ and he refuses to ‘belittle this human tendency’.

Wittgenstein’s assessment demands the question: ‘But *why* do we participate in this “nonsense”?’ and an evolutionary story like that favoured by Ruse begins the answer. But another question beckons for both Wittgenstein and Ruse: ‘Surely to see nonsense for what it is requires, on pain of irrationality, its rejection?’ The argument of this last section explores one way of replying ‘Not necessarily.’ The question of what we ought to *do*, once we have come to see that our moral discourse is a philosophically indefensible illusion, is a practical question. A neglected answer is that the discourse may be maintained, accepted, but not believed – that it may have the role of a fiction. There is nothing irrational about fictions (so long as we don’t believe them); there is nothing irrational about our allowing them to influence our emotions and decisions, or even thinking them of immense importance. Given that the widespread tendency to resist a moral error theory – to think of it as a *dangerous* doctrine – surely does not arise from the manifest plausibility or lucidity of moral concepts, but rather from a fear of what might *happen* if we abolished them, it seems to me quite likely that the practically optimal course, and therefore the rational course – both for society considered collectively, and for the individual – will be to keep these concepts alive.

Notes

¹ I believe that Mackie, who gave us the term ‘error theory’, would have disagreed with little in Ruse’s overall project. Although the evolutionary aspect of Mackie’s theory is underdeveloped, there is little doubt that he saw morality as an essentially biological phenomenon (see Mackie 1977: 113).

² This is not to say that evolution has favoured cooperation with *anyone* in *any* circumstances. Of course not. Nor do I maintain that morality can be understood entirely in terms of cooperative actions (and sentiments favouring those actions) – attitudes towards various *self-regarding* actions have quite possibly also been selected for. Also, although the disposition to see certain activities and traits as ‘intrinsically required’ naturally developed in relation to *cooperative* tendencies, there is no reason why cultural pressures might not come to transfer that sense of requirement to other types of action (e.g., in Catholic priests, to celibacy); thus there will be significant cross-cultural differences among moral systems. What they share, at a minimum, is a sense that some actions ‘must be (not) done, regardless of the performer’s ends’, and these required actions will most *probably* attach to cooperative behaviour. These

are important and complex qualifications, but they are not the subject of the present paper, where I keep things simple for brevity.

³ I say 'untrue' rather than 'false', since the correct conclusion might be that the abstract singular term 'moral obligatoriness' fails to refer to any property at all (as opposed to referring to a property which nothing actually has), in which case one might, for familiar Strawsonian reasons, hold that 'Moral obligatoriness is had by ϕ ' is *neither true nor false* (like 'The present king of France is wise'). Since that sentence, arguably, expresses the same proposition (if any) as ' ϕ is morally obligatory', the latter too would be neither true nor false.

⁴ Hume (*Treatise*, Book III, part I, section 1) writes: 'Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason'. Many modern ethicists have agreed with him. Michael Smith, though disagreeing with Hume's apparently noncognitivist conclusion, argues in detail for the thesis that one who makes a moral judgment (assuming she is practically rational) feels *prima facie* motivation. See Smith (1994).

⁵ I must say, in fairness to Richards, that he does *not* think that the mere fact that we have, as a product of natural selection, a disposition to favour altruism entails that we ought to be altruistic. He notes that we also have evolved aggressive tendencies, but he doesn't think it follows that we ought to act on them. See Richards (1986: 288, 342). However, I must admit that I do not properly understand Richards' attempt to argue for a principled distinction on this point.

⁶ It is important to stress that the sense of 'disposition' under discussion is specific: an inherited trait that regulates the formation of certain attitudes when the agent is exposed to certain environmental cues at a certain point in development. Thus when I claim that Jack 'has the disposition', this is a claim about his genetic package; it does not follow that there are any environmental stimuli that Jack could encounter *now* that would result in his forming the attitudes in question.

⁷ In particular, I have never understood why William Frankena's sensible 1939 article did not put an end to the whole business.

⁸ Compare the kind of 'non-natural' property that Moore thought is the referent of 'good'. If we had a well-confirmed theory that explained all relevant phenomena by appeal only to our *making judgments* that such non-natural properties exist, then Ockham's Razor should serve to establish an error theory – for in order for those judgments to be *true* we would be required to posit some extra kind of entity in the world (i.e., non-natural properties), but this additional ontology would not explain anything that was not explained by the theory that appealed only to (untrue) judgments.

⁹ See my 'Rational Fear of Monsters', *British Journal of Aesthetics* (April, 2000) for further discussion along these lines.

¹⁰ These thoughts are developed at further length in my 'Moral Fictionalism' (forthcoming).

References

- Anscombe, G.E.M.: 1958, 'Modern Moral Philosophy', *Philosophy* **33**, 1–19.
- Bentham, J.: 1843, 'Essay on Logic', in *Collected Works*, Volume VIII, William Tait, Edinburgh.
- Boyd, R.: 1988, 'How to be a Moral Realist', in G. Sayre-McCord (ed.), *Essays in Moral Realism*, Cornell University Press, Ithaca.
- Brink, D.: 1984, *Moral Realism and the Foundations of Ethics*, Cambridge University Press, Cambridge.

- Campbell, J.: 1993, 'A Simple View of Colour', in J. Haldane and C. Wright (eds), *Reality, Representation and Projection*, Oxford University Press, New York.
- Foot, P.: 1972, 'Morality as a System of Hypothetical Imperatives', *Philosophical Review* **81**, 305–316.
- Frankena, W.: 1939, 'The Naturalistic Fallacy', *Mind* **48**, 464–477.
- Frankena, W.: 1973, *Ethics*, Prentice Hall, Englewood Cliffs.
- Harman, G.: 1975, 'Reasons', *Critica* **7**, 3–13.
- Harman, G.: 1977, *The Nature of Morality*, Oxford University Press, New York.
- Hughes, W.: 1986, 'Richard's Defence of Evolutionary Ethics', *Biology and Philosophy* **1**, 306–315.
- Hume, D.: 1978, *A Treatise of Human Nature*, Clarendon Press, Oxford.
- Hume, D.: 1983, *Enquiry Concerning the Principles of Morals*, Hackett Publishing Company, Indianapolis.
- Johnston, M.: 1992, 'How to Speak of the Colors', *Philosophical Studies* **68**, 221–263.
- Joyce, R.: 2000, 'Rational Fear of Monsters', *The British Journal of Aesthetics* **40**, 209–224.
- Kant, I.: 1993, *Groundwork to the Metaphysics of Morals*, translated by H.J. Paton, Routledge, London.
- Mackie, J.: 1977, *Ethics: Inventing Right and Wrong*, Penguin Books, New York.
- McDowell, J.: 1985, 'Values and Secondary Qualities', in T. Honderich (ed.), *Morality and Objectivity*, Routledge and Kegan Paul, London.
- Moore, G.E.: 1903, *Principia Ethica*, Cambridge University Press, Cambridge.
- Putnam, H.: 1981, *Reason, Truth and History*, Cambridge University Press, Cambridge.
- Richards, R.J.: 1986, 'A Defence of Evolutionary Ethics', *Biology and Philosophy* **1**, 265–293.
- Rottschaefer, W.A.: 1998, *The Biology and Psychology of Moral Agency*, Cambridge University Press, Cambridge.
- Rottschaefer, W.A. and Martinsen, D.: 1990, 'Really Taking Darwin Seriously: An Alternative to Michael Ruse's Darwinian Metaethics', *Biology and Philosophy* **5**, 149–173.
- Ruse, M.: 1986a, *Taking Darwin Seriously*, Basil Blackwell, Oxford.
- Ruse, M.: 1986b, 'Evolutionary Ethics: A Phoenix Risen', *Zygon* **21**, 95–112.
- Smith, M.: 1993, 'Objectivity and Moral Realism: On the Significance of the Phenomenology of Moral Experience', in J. Haldane and C. Wright (eds), *Reality Representation & Projection*, Oxford University Press, New York.
- Smith, M.: 1994, *The Moral Problem*, Blackwell, Oxford.
- Wittgenstein, L.: 1965, 'Lecture on Ethics', *Philosophical Review* **74**, 3–12.
- Woolcock, P.: 1993, 'Ruse's Darwinian Meta-Ethics: A Critique', *Biology and Philosophy* **8**, 423–439.