Preface

When Peter Momtchiloff invited us to edit *The Oxford Handbook of Contemporary Philosophy* we sat down (over a glass of wine, truth be known) and asked ourselves how best to produce a volume that, while not being an encyclopedia, was not a handbook of one or another area of philosophy. We wanted a volume that would give readers a sense of the range and excitement of contemporary analytic philosophy (excluding formal logic) and would inform them of some of the most interesting recent developments, while being something they could hold in one hand or maybe cradle in two.

We also wanted a volume that would be a contribution to the subject. With this in mind, we invited our contributors to take the opportunity to set agendas for future discussions of the subject matters of their chapters. They were asked to produce chapters that gave a good sense of the philosophical geography of their assigned topic, but we gave them maximum flexibility in how to structure their chapters and made it clear that they were free to focus the discussion on the issues they judged to be most central and to express their own opinions. We were looking not for a mini-encyclopedia but, if you like, for a series of very high-quality opinion pieces. We were delighted with the response. Reading the chapters as they came in was an education in the contemporary philosophical scene for both of us.

Although we gave our contributors maximum flexibility, we were intrusive when it came to the topics within the various parts (moral philosophy, social and political philosophy, philosophy of mind and action, philosophy of language, metaphysics, epistemology, and philosophy of the sciences). For each part we made a judgement concerning the topics of most interest and fertility, and of course drew on our knowledge of who was working on what. For example, in the philosophy of the sciences it seemed to us that realism, laws, physics, and biology were four topics that stood out for inclusion, and we were delighted to attract four major players on those topics as contributors. Similar remarks apply to the other parts.

An example of where we drew on our knowledge of who was working on what is the chapter by John Doris and Stephen Stich, 'As a Matter of Fact: Empirical Perspectives on Ethics'. We had heard versions of the challenging ideas in this chapter as presentations. But in fact most of the invitations to our contributors were prompted in one way or another by personal acquaintance with their work. There are also a number of chapters that we knew were in someone's head and that what was needed to make the highly desirable transfer from head to page was the right invitation.

There are topics we wish we could have included but could not find room for—or the right contributor for; and, of course, other editors would have made different choices. That's life.

Producing this volume has been a lot of work—perhaps rather more than we had expected. We are very grateful to our contributors for their contributions and in some cases their extraordinary patience, and to Peter Momtchiloff and Laurien Berkeley of Oxford University Press.

F. J. and M. S.

CHAPTER 5

AS A MATTER OF FACT: EMPIRICAL PERSPECTIVES ON ETHICS

JOHN M. DORIS AND
STEPHEN P. STICH

> Too many moral philosophers and commentators on moral philosophy . . . have been content to invent their psychology or anthropology from scratch. . . .
>
> S. Darwall, A. Gibbard, and P. Railton (1997: 34–5)

1. Introduction

Regarding the assessment of Darwall and colleagues, we couldn't agree more: far too many moral philosophers have been content to *invent* the psychology or anthropology on which their theories depend, advancing or disputing empirical

claims with little concern for empirical evidence. We also believe—and we expect Darwall, Gibbard, and Railton would agree—that this empirical complacency has impeded progress in ethical theory and discouraged investigators in the biological, behavioural, and social sciences from undertaking philosophically informed research on ethical issues.

We realize that some moral philosophers have taken there to be good reasons for shunning empirical inquiry. For much of the twentieth century, many working in analytic ethics—variously inspired by Hume's (1978: 469) pithy injunction against inferring *ought* from *is* and the seductive mysteries of Moore's (1903, esp. 10–17) 'Open Question Argument'—maintained that descriptive considerations of the sort adduced in the natural and social sciences cannot constrain ethical reflection without vitiating its prescriptive or normative character (e.g. Stevenson 1944: 108–10; R. M. Hare 1952: 79–93). The plausibility of such claims is both debated and debatable, but it is clear that they have helped engender suspicion regarding 'naturalism' in ethics, which we understand, broadly, as the view that *ethical theorizing should be an (in part) a posteriori inquiry richly informed by relevant empirical considerations*.[1] Relatedly, this anti-naturalist suspicion enables disciplinary xenophobia in philosophical ethics, a reluctance to engage research beyond the philosophical literature. The methodology we advocate here—a resolutely naturalistic approach to ethical theory squarely engaging the relevant biological, behavioural, and social sciences—flouts both of these anxieties.

Perhaps those lacking our equanimity suspect that approaches of the sort we endorse fail to heed Stevenson's (1963: 13) advice that 'Ethics must not be psychology', and thereby lapse into a noxious 'scientism' or 'eliminativism'. Notoriously, Quine (1969: 75) advocated eliminativism in his rendering of naturalized epistemology, urging philosophical 'surrender of the epistemological burden to psychology'. Quine was sharply rebuked for slighting the normative character of epistemology (e.g. Kim 1988; Stich 1993*a*), but we are not suggesting, in a rambunctiously Quinean spirit, 'surrender of the *ethical* burden to psychology'. And so far as we know, neither is anyone else. Ethics must not—indeed cannot—*be* psychology, but it does not follow that ethics should *ignore* psychology.

The most obvious, and most compelling, motivation for our perspective is simply this: It is not possible to step far into the ethics literature without stubbing one's toe on empirical claims. The thought that moral philosophy can proceed unencumbered by facts seems to us an unlikely one: there are just too many places where answers to important ethical questions require—and have very often presupposed—answers to empirical questions.

A small but growing number of philosophers, ourselves included, have become convinced that answers to these empirical questions should be informed by systematic

[1] Compare Railton's (1989: 155–6) 'methodological naturalism'.

empirical research.[2] This is not to say that relevant information is easy to come by: the science is not always packaged in forms that are easy on the philosophical digestion. As Darwall *et al.* (1997: 47 ff.) caution, one won't often find 'a well-developed literature in the social sciences simply awaiting philosophical discovery and exploitation'. Still, we are more optimistic than Darwall and colleagues about the help philosophers can expect from empirical literatures: science has produced much experimental and theoretical work that appears importantly relevant to ongoing debates in ethical theory, and some moral philosophers have lately begun to pursue empirical investigations. To explore the issues fully requires far more space than is available here; we must content ourselves with developing a few rather programmatic examples of how an empirically sensitive philosophical ethics might proceed.

Our point is not that reference to empirical literatures can be expected, by itself, to resolve debates in moral theory. Rather, we hope to convince the reader that these literatures are often deeply relevant to important debates, and it is therefore intellectually irresponsible to ignore them. Sometimes empirical findings seem to contradict what particular disputing parties assert or presuppose, while in other cases, they appear to reconfigure the philosophical topography, revealing that certain lines of argument must traverse empirically difficult terrain. Often, philosophers who follow these challenging routes will be forced to make additional empirical conjectures, and these conjectures, in their turn, must be subject to empirical scrutiny. The upshot, we conclude, is that an intellectually responsible philosophical ethics is one that continuously engages the relevant empirical literature.

2. Character

In the second half of the twentieth century the 'ethics of virtue' became an increasingly popular alternative to the Kantian and utilitarian theories that had for some time dominated normative ethics. In contrast to Kantianism and utilitarianism, which despite marked differences share an emphasis on identifying morally obligatory actions, virtue-centred approaches emphasize the psychological constitution, or character, of actors. The central question for virtue ethics, so the slogan goes, is not what sort of action to do, but what sort of person to be.[3] As Bernard Williams

[2] See Gibbard (1990: 58–61); Flanagan (1991); Goldman (1993); Johnson (1993); Stich (1993*b*); Railton (1995); Blackburn (1998: 36–7); Bok (1996); Doris (1996, 1998, 2002); Becker (1998); Campbell (1999); Harman (1999, 2000); Merritt (1999, 2000); Doris and Stich (2001); Woolfolk and Doris (2002).

[3] The notion that character is evaluatively independent of or prior to action is sometimes thought to be the distinctive emphasis of virtue ethics (see Louden 1984: 229; Watson 1990: 451–2). But this is

(1985: 1) has eloquently reminded us, the 'aims of moral philosophy, and any hopes it may have of being worth serious attention, are bound up with the fate of Socrates' question' How should one live?, and it has seemed to many philosophers, not least due to Williams's influence, that any prospects for a satisfying answer rest with the ethics of character. Allegedly, if ethical reflection is to help people understand and improve themselves and their relations to others, it must be reflection focused on the condition and cultivation of character (see Williams 1993: 91–5).

Virtue ethics, especially in the Aristotelian guises that dominate the field, typically presupposes a distinctive account of human psychology. Nussbaum (1999: 170), although she insists that the moniker 'virtue ethics' has been used to tag such a variety of projects that it represents a 'misleading category', observes that approaches so titled are concerned with the 'settled patterns of motive, emotion, and reasoning that lead us to call someone a person of a certain sort (courageous, generous, moderate, just, etc.)'. If this is a fair characterization—and we think it is—then virtue ethics is marked by a particular interest in moral psychology, an interest in the cognitive, affective, and emotional patterns that are associated with the attribution of character traits.[4] This interest looks to be an empirical interest, and it's natural to ask how successfully virtue ethics addresses it.

The central empirical issue concerns, to borrow Nussbaum's phrase, 'settled patterns' of functioning. According to Aristotle, genuinely virtuous action proceeds from 'firm and unchangeable character' rather than from transient motives (1984: $1105^{a}28$–$^{b}1$); while the good person may suffer misfortune that impairs his activities and diminishes happiness, he 'will never (*oudepote*) do the acts that are hateful and mean' (1984: $1100^{b}32$–4; cf. $1128^{b}29$; cf. Cooper 1999: 299 ff.).[5] In an influential contemporary exposition, McDowell (1978: 26–7) argued that considerations favouring vicious behaviour are 'silenced' in the virtuous person; although such an individual may recognize inducements to vice, she will not count them as reasons for action. As we understand the tradition, virtues are supposed to be robust traits; if a person has a robust trait, she can be confidently (although perhaps not with absolute certainty) expected to display trait-relevant behaviour across a wide variety of

not plausibly understood to mean that virtue ethics is indifferent regarding questions of what to do; the question of conduct should be of substantial importance on both virtue- and action-centred approaches (see Sher 1998: 15–17).

[4] Nussbaum (1999: 170) observes that Kantian and utilitarian approaches may share virtue ethics' interest in character. Space prohibits discussion, but if Nussbaum were right, our argument would have more sweeping implications than we contemplate here.

[5] In Aristotle's view, the virtues are *hexeis* (1984: $1106^{a}10$–12), and a *hexis* is a disposition that is 'permanent and hard to change' (1984: $8^{b}25$–$9^{a}9$). This feature of Aristotle's account is emphasized by commentators: Sherman (1989: 1) says that for Aristotle (as well as for us) character traits explain why 'someone can be *counted on* to act in certain ways' (cf. Woods 1986: 149; Annas 1993: 51; Audi 1995: 451; Cooper 1999: 238).

trait-relevant situations, even where some or all of these situations are not optimally conducive to such behaviour (Doris 2002: 18).[6]

Additionally, some philosophers have supposed that character will be evaluatively integrated—traits with associated evaluative valences are expected to co-occur in personality (see Doris 2002: 22; Flanagan 1991: 283–90). As Aristotle (1984: $1144^{b}30$–$1145^{a}2$; cf. Irwin 1988: 66–71) has it, the virtues are inseparable; given the qualities of practical reason sufficient for the possession of one virtue, one can expect to find the qualities of practical reason sufficient for them all.

While understandings of character and personality akin to those just described have been hotly contested in psychology departments at least since the critiques of Vernon (1964), Mischel (1968), and Peterson (1968), moral philosophers have not been especially quick in taking the matter up. Flanagan's (1991) careful discussion broached the issue in contemporary analytic ethics, while Doris (1998, 2002) and Harman (1999, 2000) have lately pressed the point less temperately: although they manifest some fraternal disagreements, Harman and Doris both insist that the conception of character presupposed by virtue ethics is empirically inadequate.

The evidence for this contention, often united under the theoretical heading of 'situationism', has been developed over a period of some seventy years, and includes some of the most striking research in the human sciences.

- Mathews and Canon (1975: 574–5) found subjects were five times more likely to help an apparently injured man who had dropped some books when ambient noise was at normal levels than when a power lawnmower was running nearby (80 per cent v. 15 per cent).
- Darley and Batson (1973: 105) report that passers-by not in a hurry were six times more likely to help an unfortunate who appeared to be in significant distress than were passers-by in a hurry (63 per cent v. 10 per cent).
- Isen and Levin (1972: 387) discovered that people who had just found a dime were twenty-two times more likely to help a woman who had dropped some papers than those who did not find a dime (88 per cent v. 4 per cent).
- Milgram (1974) found that subjects would repeatedly 'punish' a screaming 'victim' with realistic (but simulated) electric shocks at the polite request of an experimenter.
- Haney *et al.* (1973) describe how college students role-playing in a simulated prison rapidly descended to *Lord of the Flies* barbarism.

There apparently exists an alarming disproportion between situational input and morally disquieting output; it takes surprisingly little to get people behaving in

[6] This follows quite a standard theme in philosophical writings on virtue and character. For example, Blum (1994: 178–80) understands compassion as a trait of character typified by an altruistic attitude of 'strength and duration', which should be 'stable and consistent' in prompting beneficent action (cf. Brandt 1970: 27; Dent 1975: 328; McDowell 1979: 331–3; Larmore 1987: 12).

morally undesirable ways. The point is not that circumstances influence behaviour, or even that seemingly good people sometimes do lousy things. No need to stop the presses for that. Rather, the telling difficulty is just how insubstantial the situational influences effecting troubling moral failures seem to be; it is not that people fall short of ideals of virtue and fortitude, but that they can be *readily* induced to *radically* fail such ideals.

The argument suggested by this difficulty can be outlined as follows: a large body of research indicates that cognition and behaviour are extraordinarily sensitive to the situations in which people are embedded. The implication is that individuals—on the altogether plausible assumption that most people will be found in a range of situations involving widely disparate cognitive and behavioural demands—are typically highly variable in their behaviour, relative to the behavioural expectations associated with familiar trait categories such as honesty, compassion, courage, and the like. But if people's behaviour were typically structured by robust traits, one would expect quite the opposite: namely, behaviour consistent with a given trait—e.g. behaviour that is appropriately and reliably honest, compassionate, or courageous—across a diversity of situations. It follows, according to the argument, that behaviour is not typically structured by the robust traits that figure centrally in virtue-theoretic moral psychology. Analogous considerations are supposed to make trouble for notions of evaluative integration; the endemic lack of uniformity in behaviour adduced from the empirical literature undermines expectations of integrated character structures.

The situationist argument has sometimes been construed by philosophers as asserting that character traits 'do not exist' (Flanagan 1991: 302; Athanassoulis 2000: 219–20; Kupperman 2001: 250), but this is a misleading formulation of the issue.[7] In so far as to deny the existence of traits is to deny the existence of persisting dispositional differences among persons, the claim that traits do not exist seems unsustainable, and the exercise of refuting such a claim idle. (Indeed, it is a claim that even psychologists with strong situationist sympathies, e.g. Mischel 1968: 8–9, seem at pains to disavow.) The real issue dividing the virtue theorist and the situationist concerns the appropriate characterization of traits, not their existence or non-existence. The situationist argument that needs to be taken seriously, and which to our mind stands unrefuted, holds that the Aristotelian conception of traits as robust dispositions—the sort which lead to trait-relevant behaviour across a wide variety of trait-relevant situations—is radically empirically undersupported. To

[7] Part of the reason for this error may be some spirited rhetoric of Harman's (e.g. the title of Harman 2000: 'The Nonexistence of Character Traits'). But Harman repeatedly offers qualifications that caution against it; he voices scepticism about the existence of 'ordinary character traits *of the sort people think there are*' (1999: 316) and 'character traits *as ordinarily conceived*' (2000: 223; our italics). This is to reject a particular conception of character traits, not to deny that character traits exist. For his part, Doris (1998: 507; 2002: 62–6) quite explicitly acknowledges the existence of traits, albeit traits with less generalized effects on behaviour than is often supposed.

put the ethical implications of this a bit aggressively, it looks as though attribution of robust traits like virtues may very well be unwarranted in most instances,[8] programmes of moral education aimed at inculcating virtues may very well be futile, and modes of ethical reflection focusing moral aspirations on the cultivation of virtue may very well be misguided.

At this point, the virtue theorist may offer one of two responses. She can accept the critics' interpretation of the empirical evidence while denying that her approach makes empirical commitments of the sort the evidence indicates is problematic. Or she can allow that her approach makes commitments in empirical psychology of the sort that would be problematic if the critics' interpretations of the evidence were sustainable, but deny that the critics have interpreted the evidence aright. The first option, we might say, is 'empirically modest' (see Doris 2002: 110–12): because such renderings make only minimal claims in empirical psychology, they are insulated from empirical threat. The second option, conversely, is 'empirically vulnerable' (see Railton 1995: 92–6): it makes empirical claims with enough substance to invite empirically motivated criticism.

We shall first discuss empirically modest rejoinders to the situationist critique. Numerous defenders of virtue ethics insist that virtue is not expected to be widely instantiated, but is found in only a few extraordinary individuals, and these writers further observe that this minimal empirical commitment is quite compatible with the disturbing, but not exceptionlessly disturbing, behaviour in experiments like Milgram's (see Athanassoulis 1999: 217–19; DePaul 1999; Kupperman 2001: 242–3). The critics are bound to concede the point, since the empirical evidence cannot show that the instantiation of virtue in actual human psychologies is impossible; no empirical evidence could secure so strong a result. But so construed, the aspirations of virtue ethics are not entirely clear; if virtue is *expected* to be rare, it is not obvious what role virtue theory could have in a (generally applicable) programme of moral education.[9] This rings a bit odd, given that moral education—construed as aiming for the development of the good character necessary for a good life—has traditionally been a distinctive emphasis in writing on virtue, from Aristotle (1984: $1099^{b}9$–32, $1103^{b}3$–26) to Bennett (1993: 11–16; cf. Williams 1985: 10). Of course, the rarity of virtue might be thought a contingent matter; given the appropriate modalities of moral education, the virtue ethicist might say, virtue can be widely inculcated. But philosophers, psychologists, and educators alike have tended to be a bit hazy regarding particulars of the requisite educational processes; theories of moral

[8] The difficulty is not limited to rival academic theories; there is a large body of empirical evidence indicating that everyday 'lay' habits of person perception seriously overestimate the impact of individual dispositional differences on behavioural outcomes. For summaries, see Jones (1990); Ross and Nisbett (1991: 119–44); Gilbert and Malone (1995); Doris (2002: 92–106).

[9] Of course, if the virtue theorist is an elitist, this need not trouble her. But while historical writers on the virtues have at times manifested elitist sympathies (Aristotle 1984: $1123^{a}6$–10, $1124^{a}17$–$^{b}32$; Hume 1975: 250–67), this is not a sensibility that is typically celebrated by contemporary philosophers.

education, and character education in particular, are typically not supported by large bodies of systematic research adducing behavioural differences corresponding to differing educational modalities (Leming 1997*a*,*b*; Hart and Killen 1999: 12; Doris 2002: 121–7).

It is tempting to put the situationist point a bit more sharply. It is true that the evidence does not show that the instantiation of virtue in actual human psychologies is impossible. But it also looks to be the case that the available systematic empirical evidence is compatible with virtue being psychologically impossible (or at least wildly improbable), and this suggests that the impossibility of virtue is an empirical possibility that has to be taken seriously. So while the evidence doesn't refute an empirically modest version of virtue ethics, it is plausibly taken to suggest that the burden of argument has importantly shifted: The advocate of virtue ethics can no longer simply assume that virtue is psychologically possible. If she can't offer compelling evidence—very preferably, more than anecdotal evidence—favouring the claim that virtue is psychologically possible, then she is in the awkward position of forwarding a view that would be undermined if an empirical claim which is not obviously false were to turn out to be true, without offering compelling reason to think that it won't turn out to be true.

Suppose the realization of virtue were acknowledged to be impossible: it might yet be insisted that talk of virtue articulates ethical ideals that are well suited—presumably better suited than alternatives, if virtue ethics is thought to have distinctive advantages—to facilitating ethically desirable conduct (see Blum 1994: 94–6). Asserting such a practical advantage for virtue ethics entails an empirical claim: reflection on the ideals of virtue can help actual people behave better. For example, it might be claimed that talk of virtue is more compelling, or has more motivational 'grip', than abstract axiological principles. We know of little systematic evidence favouring such claims, and we are unsure of what sort of experimental designs are fit to secure them, but the only point we need to insist on is that even this empirically modest rendering of virtue ethics may bear contentious empirical commitments. If virtue ethics is alleged to have practical implications, it cannot avoid empirical assertions regarding the cognitive and motivational equipment with which people navigate their moral world.

Even without an answer to such practical questions, it might be thought that virtue ethics is fit to address familiar conceptual problems in philosophical ethics, such as rendering an account of right action. In Hursthouse's (1999: 28; cf. 49–51) account of virtue ethics, 'An action is right iff it is what a virtuous agent would characteristically (i.e., acting in character) do in the circumstances.' Hursthouse (1999: 123–6, 136, 140) further insists that an action does not count as 'morally motivated' simply by dint of being the sort of thing a virtuous person does, done for reasons of the sort the virtuous person does it for; it must proceed 'from virtue', that is, 'from a settled state of good character'. If this requirement is juxtaposed with the observation that the relevant states of character are extremely rare, as an

empirically modest rendering of virtue ethics maintains, we apparently get the result that 'morally motivated' actions are also extremely rare (a virtue-theoretic result, interestingly, with which Kant would have agreed). This need not trouble Hursthouse (1999: 141–60); she seems to allow that very often—perhaps always—one sees only approximations of moral motivation. It does trouble us. We think that less than virtuous people, even smashingly less than virtuous people, sometimes do the right thing for the right reasons, and these actions are fit to be honoured as 'morally motivated'. It may not happen as often as one would like, but morally motivated conduct seems to happen rather more frequently than one chances on perfect virtue. Oskar Schindler, the philandering war profiteer who rescued thousands of Jews from the Nazis, is a famous example of the two notions coming apart (see Kenneally 1982), but with a little attention to the history books, we can surely adduce many more. The burden of proof, it seems to us, is on those asserting that such widely revered actions are not morally motivated.

There are also serious questions about the competitive advantages enjoyed by empirically modest virtue ethics. It has seemed to many that a chief attraction of character-based approaches is the promise of a lifelike moral psychology—a less wooden depiction of moral affect, cognition, motivation, and education than that offered by competing approaches such as Kantianism and utilitarianism (Flanagan 1991: 182; Hursthouse 1999: 119–20). Proponents of virtue ethics, perhaps most prominently MacIntyre (1984) and Williams (1985, 1993), link their approach—as Anscombe (1958: 4–5) did in a paper widely regarded as the call to arms for contemporary virtue ethics—to prospects for more psychological realism and texture. We submit that this is where a large measure of virtue ethics' appeal has lain; if virtue ethicists had tended to describe their psychological project along the lines just imagined, as deploying a moral psychology only tenuously related to the contours of actual human psychologies, we rather doubt that the view would now be sweeping the field.

We contend that for virtue ethics to retain its competitive advantage in moral psychology it must court empirical danger by making empirical claims with enough substance to be seriously tested by the empirical evidence from psychology. For instance, the virtue theorist may insist that while perfect virtue is rare indeed, robust traits approximating perfect virtue—reliable courage, temperance, and the rest—may be widely inculcated, and perhaps similarly for robust vices—reliable cowardice, profligacy, and so on.[10] To defend such a position, the virtue theorist must somehow discredit the critic's empirical evidence. Various arguments might be thought to secure such a result: (i) The situationist experiments might be methodologically flawed; problems in experimental design or data analysis, for example, might undermine the results. (ii) The experiments might fail standards of ecological

[10] There is some question as to whether vices are expected to be robust in the way virtues are, but some philosophers seem to think so: Hill (1991: 130–2) apparently believes that calling someone weak-willed marks characteristic patterns of behaviour.

validity; the experimental contexts might be so distant from natural contexts as to preclude generalizations to the 'real world'. (iii) General conclusions from the experiments might be prohibited by limited samples; in particular, there appears to be a dearth of longitudinal behavioural studies that would help assess the role of character traits 'over the long haul'. (iv) The experiments may be conceptually irrelevant; for example, the conceptions of particular traits operationalized in the empirical work may not correspond to the related conceptions figuring in virtue ethics.

The thing to notice straight away is that motivating contentions like the four above require evaluating a great deal of psychological research; making a charge stick to one experiment or two, when there are hundreds, if not thousands, of relevant studies, is unlikely to effect a satisfying resolution of the controversy. The onus, of course, falls on both sides: just as undermining arguments directed at single experiments are of limited comfort to the virtue theorist, demonstrating the philosophical relevance of a lone study is not enough to make the critics' day. Newspaper science reporting notwithstanding, in science there is seldom, or never, a single decisive experiment or, for that matter, a decisive experimental failure. General conclusions about social science can legitimately be drawn only from encountering, in full detail, a body of research, and adducing patterns or trends. Doris (2002) has recently attempted to approximate this methodological standard in a book-length study, and he there concludes that major trends in empirical work support conclusions in the neighbourhood indicated by the more programmatic treatments of Doris (1998) and Harman (1999, 2000). Whether or not one is drawn to this conclusion, we think it clear that the most profitable discussion of the empirical literature will proceed with detailed discussion of the relevant empirical work. If an empirically vulnerable virtue ethics is to be shown empirically defensible, defenders must provide much fuller consideration of the psychology. To our knowledge, extant defences of virtue ethics in the face of empirical attack do not approximate the required breadth and depth.[11] Hopefully, future discussions will rectify this situation, to the edification of defenders and critics alike.

3. Moral Motivation

Suppose a person believes that she ought to do something: donate blood to the Red Cross, say, or send a significant contribution to an international relief agency. Does it follow that she will be moved actually to act on this belief? Ethical theorists use

[11] For example, Kupperman (2001) refers to nine items in the empirical literature in responding to Harman, and Athanassoulis (2000), three.

internalism to mark an important cluster of answers to this question, answers maintaining that the motivation to act on a moral judgement is a necessary or intrinsic concomitant of the judgement itself, or that the relevant motivation is inevitably generated by the very same mental faculty that produces the judgement.[12] One familiar version of internalism is broadly Kantian, emphasizing the role of rationality in ethics. As Deigh (1999: 289) characterizes the position, 'reason is both the pilot and the engine of moral agency. It not only guides one toward actions in conformity with one's duty, but it also produces the desire to do one's duty and can invest that desire with enough strength to overrule conflicting impulses of appetite and passion.' A notorious difficulty for internalism is suggested by Hume's (1975: 282–4) 'sensible knave', a person who recognizes that the unjust and dishonest acts he contemplates are wrong, but is completely unmoved by this realization. More recent writers (e.g. Nichols 2002) have suggested that the sensible knave (or, as philosophers often call him, 'the amoralist') is more than a philosophical fiction, since clinical psychologists and other mental health professionals have for some time noted the existence of sociopaths or psychopaths, who appear to *know* the difference between right and wrong but quite generally lack motivation to *do* what is right. If this understanding of the psychopath's moral psychology is accurate, internalism looks to be suffering empirical embarrassment.[13]

Internalists have adopted two quite different responses to this challenge, one conceptual and the other empirical. The first relies on conceptual analysis to argue that a person couldn't really believe that an act is wrong if he has no motivation to avoid performing it. For example, Michael Smith claims it is 'a conceptual truth that agents who make moral judgements are motivated accordingly, at least absent weakness of the will and the like' (Smith 1994: 66). Philosophers who adopt this strategy recognize that imaginary knaves and real psychopaths may *say* that something is 'morally required' or 'morally wrong' and that they may be expressing a judgement that they sincerely accept. But if psychopaths are not motivated in the appropriate way, their words do not mean what non-psychopaths mean by these words and the concepts they express with these words are not the ordinary moral concepts that non-psychopaths use. Therefore psychopaths 'do not *really* make moral judgements at all' (Smith 1994: 67).

This strategy only works if ordinary moral concepts require that people who *really* make moral judgements have the appropriate sort of motivation. But there is

[12] A stipulation: We refer to views in the neighbourhood of what Darwall (1983: 54) calls 'judgement internalism', the thesis that it is 'a necessary condition of a genuine instance of a certain sort of judgment that the person making the judgment be disposed to act in a way appropriate to it'. Space limitations force us to ignore myriad complications; for more detailed discussion, see Svavarsdóttir (1999).

[13] There is august precedent for supposing that the internalism debate has empirical elements. In his classic discussion, Frankena (1976: 73) observed that progress here requires reference to 'the psychology of human motivation'—'The battle, if war there be, cannot be contained; its field is the whole human world'. We hope that Frankena would have appreciated our way of joining the fight.

considerable disagreement in cognitive science about whether and how concepts are structured, and about how we are to determine when something is built into or entailed by a concept (Margolis and Laurence 1999). Indeed, one widely discussed approach maintains that concepts have no semantically relevant internal structure to be analysed—thus there are no conceptual entailments (Fodor 1998). Obviously, internalists who appeal to conceptual analysis must reject this account, and in so doing they must take a stand in the broadly empirical debate about the nature of concepts.

Smith is one moral theorist who has taken such a stand. Following Lewis (1970, 1972), Jackson (1994), and others, Smith proposes that a concept can be analysed by specifying the 'maximal consistent set of platitudes' in which the concept is invoked; it is by 'coming to treat those platitudes as platitudinous', Smith (1994: 31) maintains, that 'we come to have mastery of that concept'. If this is correct, the conceptual analysis defence of internalism requires that the maximally consistent set of platitudes invoking the notion of a moral judgement includes a claim to the effect that 'agents who make moral judgements are motivated accordingly'. Once again, this is an empirical claim. Smith appeals to his own intuitions in its support, but it is of course rather likely that opponents of internalism do not share Smith's intuitions, and it is difficult to say whose intuitions should trump.

In the interests of developing a non-partisan analysis, Nichols (2002) has been running a series of experiments in which philosophically unsophisticated undergraduates are presented with questions like these:

> John is a psychopathic criminal. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. John has hurt, and indeed killed, other people when he has wanted to steal their money. He says that he knows that hurting others is wrong, but that he just doesn't care if he does things that are wrong. Does John really understand that hurting others is morally wrong?
>
> Bill is a mathematician. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. Nonetheless, Bill never hurts other people simply because he thinks that it is irrational to hurt others. He thinks that any rational person would be like him and not hurt other people. Does Bill really understand that hurting others is morally wrong? (Nichols 2004: 74)

Nichols's preliminary results are exactly the opposite of what Smith would have one expect. An overwhelming majority of subjects maintained that John, the psychopath, did understand that hurting others is morally wrong, while a slight majority maintained that Bill, the rational mathematician, did not. The implication seems to be that the subjects' concept of moral judgement does not typically include a 'motivational platitude'. These results do not, of course, constitute a decisive refutation of Smith's conceptual analysis, since Smith can reply that responses like those Nichols reports would not be part of the maximally consistent set of platitudes that people would endorse after due reflection. But this too is an empirical claim; if Smith is to offer a compelling defence of it he should—with our enthusiastic encouragement—adduce some systematic empirical evidence.

A second internalist strategy for dealing with the problem posed by the amoralist is empirical: even if amoralists are conceptually possible, the internalist may insist, their existence is psychologically impossible. As a matter of psychological fact, this argument goes, people's moral judgements are accompanied by the appropriate sort of motivation.[14] A Kantian elaboration of this idea, on which we will focus, maintains that people's moral judgements are accompanied by the appropriate sort of motivation *unless their rational faculties are impaired*. (We'll shortly see that much turns on the fate of the italicized clause.) Recent papers by Roskies (2003) and Nichols (2002) set out important challenges to this strategy.

Roskies' argument relies on Damasio and colleagues' work with patients suffering injuries to the ventromedial (VM) cortex (Damasio *et al*. 1990; Saver and Damasio 1991; Bechara *et al*. 2000). On a wide range of standard psychological tests, including tests for intelligence and reasoning abilities, these patients appear quite normal. They also do as well as normal subjects on Kohlberg's tests of *moral* reasoning, and when presented with hypothetical situations they offer moral judgements that concur with those of normal subjects. However, these patients appear to have great difficulty acting in accordance with those judgements. As a result, although they often led exemplary lives prior to their injury, their post-trauma social lives are a shambles. They disregard social conventions, make disastrous business and personal decisions, and often engage in anti-social behaviour. Accordingly, Damasio and his colleagues describe the VM patients' condition as 'acquired sociopathy' (Saver and Damasio 1991).

Roskies maintains that VM patients do not act on their moral judgements because they suffer a *motivational* deficit. Moreover, the evidence indicates that these individuals do not have a *general* difficulty in acting on evaluative judgements; rather, Roskies (2003) maintains, action with respect to moral and social evaluation is differentially impaired. In addition to the behavioural evidence, this interpretation is supported by the anomalous pattern of skin-conductance responses (SCRs) that VM patients display.[15] Normal individuals produce an SCR when presented with emotionally charged or value-laden stimuli, while VM patients typically do not produce SCRs in response to such stimuli. SCRs are not entirely lacking in VM patients, however. SCRs are produced when VM patients are surprised or startled, for example, demonstrating that the physiological basis for these responses is intact. In addition, their presence is reliably correlated with cases in which patients' actions are consistent with their judgements about what to do, and their absence is reliably correlated with cases in which patients fail to act in accordance with their judgements. Thus, Roskies contends, the SCR is a reliable indicator of motivation.

[14] We prescind from questions as to whether the motivation need be overriding, although we suspect formulations not requiring overridingness are more plausible.

[15] SCR is a measure of physiological arousal, which is also sometimes called galvanic skin response, or GSR.

So the fact that VM patients, unlike normal subjects, do not exhibit SCRs in response to morally charged stimuli suggests that their failure to act in morally charged situations results from a motivational deficit.

On the face of it, acquired sociopathy confounds internalists maintaining that the moral judgements of rational people are, as a matter of psychological fact, always accompanied by appropriate motivation.[16] Testing indicates that the general reasoning abilities of these patients are not impaired, and even their moral reasoning seems to be quite normal. So none of the empirical evidence suggests the presence of a cognitive disability. An internalist might insist that these post-injury judgements are not *genuine* instances of moral judgements because VM patients no longer know the standard meaning of the moral words they use. But unless it is supported by an appeal to a conceptual analysis of the sort we criticized earlier, this is a rather implausible move; as Roskies notes, all tests of VM patients indicate that their language, their declarative knowledge structures, and their cognitive functioning are intact. There are, of course, many questions about acquired sociopathy that remain unanswered and much work is yet to be done. However these questions get answered, the literature on VM patients is one that moral philosophers embroiled in the internalism debate would be ill advised to ignore; once again, the outcome of a debate in ethical theory looks to be contingent on empirical issues.

The same point holds for other work on anti-social behaviour. Drawing on Blair's (1995) studies of psychopathic murderers imprisoned in Great Britain, Nichols (2002) has recently argued that the phenomenon of psychopathy poses a deep and complex challenge for internalism. Again, the general difficulty is that psychopaths seem to be living instantiations of Hume's sensible knave: although they appear to be rational and can be quite intelligent, psychopaths are manipulative, remorseless, and devoid of other-regarding concern. While psychopaths sometimes acknowledge that their treatment of other people is wrong, they are quite indifferent about the harm that they have caused; they seem to have no motivation to avoid hurting others (R. D. Hare 1993).

Blair's (1995) evidence complicates this familiar story. He found that psychopaths exhibit surprising deficits on various tasks where subjects are presented with descriptions of 'moral' transgressions like a child hitting another child and 'conventional' transgressions like a child leaving the classroom without the teacher's permission. From early childhood, normal children distinguish moral from conventional transgressions on a number of dimensions: they view moral transgressions as more serious, they explain why the acts are wrong by appeal to different factors (harm and fairness for moral transgressions, social acceptability for conventional transgressions), and they understand conventional transgressions, unlike moral transgressions, to be dependent on authority (Turiel *et al.* 1987; Nucci 1986).

[16] Roskies herself does not offer acquired sociopathy as a counter-example to the Kantian version of empirical internalism, but we believe the evidence *is* in tension with the Kantian view we describe.

For example, presented with a hypothetical case where a teacher says there is no rule about leaving the classroom without permission, children think it is OK to leave without permission. But presented with a hypothetical where a teacher says there is no rule against hitting other children, children do not judge that hitting is acceptable. Blair has shown that while autistic children, children with Down syndrome, and a control group of incarcerated non-psychopath murderers have relatively little trouble in drawing the moral–conventional distinction and classifying cases along these lines, incarcerated psychopaths are unable to do so.

This inability might be evidence for the hypothesis that psychopaths have a reasoning deficit, and therefore do not pose a problem for internalists who maintain that a properly functioning reasoning faculty reliably generates some motivation to do what one believes one ought to do. But, as Nichols (2002) has pointed out, the issue cannot be so easily resolved, because psychopaths have also been shown to have *affective* responses that are quite different from those of normal subjects. When shown distressing stimuli (like slides of people with dreadful injuries) and threatening stimuli (like slides of an angry man wielding a weapon), normal subjects exhibit much the same suite of physiological responses. Psychopaths, by contrast, exhibit normal physiological responses to threatening stimuli, but abnormally low physiological responses to distressing stimuli (Blair *et al.* 1997). Thus, Nichols argues, it may well be that the psychopath's deficit is not an abnormal reasoning system, but an abnormal affect system, and it is these affective abnormalities, rather than any rational disabilities, that are implicated in psychopaths' failure to draw the moral–conventional distinction.[17] If his interpretation is correct, it looks as though the existence of psychopaths does undermine the Kantian internalist's empirical generalization: contra the Kantian, there exists a substantial class of individuals *without rational disabilities* who are not motivated by their moral judgements.

We are sympathetic to Nichols's account, but as in the case of VM patients, the internalist is free to insist that a fuller understanding of psychopathy will reveal that the syndrome does indeed involve rational disabilities. Resolving this debate will require conceptual work on how to draw the boundary between reason and affect, and on what counts as an abnormality in each of these domains. But it will also require much more empirical work aimed at understanding exactly how psychopaths and non-psychopaths differ. The internalist—or at least the Kantian internalist—who wishes to diffuse the difficulty posed by psychopathy must proffer an empirically tenable account of the psychopath's cognitive architecture that locates the posited rational disability. We doubt that such an account is forthcoming. But—to instantiate once more our take-home message—our present point is that if internalists are to develop such an account, they must engage the empirical literature.

[17] Here Nichols offers support for the 'sentimentalist' tradition, which maintains that emotions (or 'sentiments') play a central role in moral judgement. For a helpful treatment of sentimentalism, see D'Arms and Jacobson (2000).

4. Moral Disagreement

Numerous contemporary philosophers, including Brandt (1959), Harman (1977: 125–36), Railton (1986*a*,*b*), and Lewis (1989), have proposed dispositional theories of moral rightness or non-moral good, which 'make matters of value depend on the affective dispositions of agents' (see Darwall *et al.* 1997: 28–9).[18] The various versions differ in detail,[19] but a rendering by Brandt is particularly instructive. According to Brandt (1959: 241–70), ethical justification is a process whereby initial judgements about particular cases and general moral principles are revised by testing these judgements against the attitudes, feelings, or emotions that would emerge under appropriately idealized circumstances. Of special importance on Brandt's (1959: 249–51, 261–4) view are what he calls 'qualified attitudes'—the attitudes people would have if they were, *inter alia*, (1) impartial, (2) fully informed about and vividly aware of the relevant facts, and (3) free from any 'abnormal' states of mind, like insanity, fatigue, or depression.[20]

As Brandt (1959: 281–4) noted, much depends on whether all people would have the same attitudes in ideal circumstances—i.e. on whether their attitudes would *converge* in ideal circumstances. If they would, then certain moral judgements—those where the idealized convergence obtains—are justified for all people, and others—those where such convergence fails to obtain—are not so justified. But if people's attitudes generally fail to converge under idealized circumstances, qualified attitude theory apparently lapses into a version of relativism, since any given moral judgement may comport with the qualified attitudes of one person, and thus be justified for him, while an incompatible judgement may comport with the attitudes of another person, and thus be justified for her.[21]

Brandt, who was a pioneer in the effort to integrate ethical theory and the social sciences, looked primarily to anthropology to help determine whether moral attitudes can be expected to converge under idealized circumstances. It is of course

[18] These views reflect a venerable tradition linking moral judgement to the affective states that people would have under idealized conditions; it extends back to Hutcheson (1738), Hume (1975, 1978), and Adam Smith (2002).

[19] A particularly important difference concerns the envisaged link between moral claims and affective reactions. Firth (1952: 317–45) and Lewis (1989) see the link as a matter of meaning, Railton (1986*b*) as a synthetic identity, and Brandt (1959: 241–70) both as a matter of justification and, more tentatively, as a matter of meaning.

[20] Brandt was a prolific and self-critical thinker, and the 1959 statement may not represent his mature views, but it well illustrates how empirical issues can impact a familiar approach to ethical theory. For a helpful survey of Brandt's career, see Rosati (2000).

[21] On some readings, qualified attitude theories may end up a version of *scepticism* if attitudes don't converge under ideal circumstances. Suppose a theory holds 'an action is morally right (or morally wrong) iff all people in ideal conditions would judge that action is morally right (or morally wrong)'. Then if convergence fails to obtain in ideal conditions, this theory entails that there are no morally right (or morally wrong) actions.

well known that anthropology includes a substantial body of work, such as the classic studies of Westermarck (1906) and Sumner (1934), detailing the radically divergent moral outlooks found in cultures around the world. But as Brandt (1959: 283–4) recognized, typical ethnographies do not support confident inferences about the convergence of attitudes under *ideal* conditions, in large measure because they often give limited guidance regarding how much of the moral disagreement can be traced to disagreement about factual matters that are not moral in nature, such as those having to do with religious or cosmological views.

With this sort of difficulty in mind, Brandt (1954) undertook his own anthropological study of Hopi people in the American southwest, and found issues for which there appeared to be serious moral disagreement between typical Hopi and white American attitudes that could not plausibly be attributed to differences in belief about non-moral facts. A notable example is the Hopi attitude towards causing animals to suffer, an attitude that might be expected to disturb many non-Hopis: '[Hopi c]hildren sometimes catch birds and make "pets" of them. They may be tied to a string, to be taken out and "played" with. This play is rough, and birds seldom survive long. [According to one informant:] "Sometimes they get tired and die. Nobody objects to this" ' (Brandt 1954: 213).

Brandt (1959: 103) made a concerted effort to determine whether this difference in moral outlook could be traced to disagreement about non-moral facts, but he could find no plausible explanation of this kind; his Hopi informants didn't believe that animals lack the capacity to feel pain, for example, nor did they believe that animals are rewarded for martyrdom in the afterlife. According to Brandt (1954: 245), the Hopi do not regard animals as unconscious or insensitive; indeed, they apparently regard animals as 'closer to the human species than does the average white man'. The best explanation of the divergent moral judgements, Brandt (1954: 245) concluded, is a 'basic difference of attitude'. Accordingly, although he cautions that the uncertainties of ethnography make confident conclusions on this point difficult, Brandt (1959: 284) argues that accounts of moral justification like his qualified attitude theory *do* end in relativism, since 'groups do sometimes make divergent appraisals when they have identical beliefs about the objects'.

Of course, the observation that persistent moral disagreement appears to problematize moral argument and justification is not unique to Brandt. While the difficulty is long familiar, contemporary philosophical discussion was spurred by Mackie's (1977: 36–8) 'argument from relativity' or, as it is called by later writers, the 'argument from disagreement' (Brink 1989: 197; Loeb 1998). Such 'radical' differences in moral judgement as are frequently observed, Mackie (1977: 36) argued, 'make it difficult to treat those judgments as apprehensions of objective truths'. As we see it, the problem is not only that moral disagreement often persists, but also that for important instances of moral disagreement—such as the treatment of animals—it is obscure what sort of considerations, be they methodological or substantive, *could* settle the issues (see Sturgeon 1988: 229). Indeed, moral disagreement might be plausibly

expected to continue even when the disputants are in methodological agreement concerning the appropriate standards for moral argument. One way of putting the point is to say that application of the same method may, for different individuals or cultures, yield divergent moral judgements that are equally acceptable by the lights of the method, even in reflective conditions that the method countenances as ideal.[22]

In contemporary ethical theory, an impressive group of philosophers are 'moral realists' (see Railton, 1986*a*,*b*; Boyd 1988; Sturgeon 1988; Brink 1989; M. Smith 1994). Adherents to a single philosophical creed often manifest doctrinal differences, and that is doubtless the case here, but it is probably fair to say that most moral realists mean to resist the argument from disagreement and reject its relativist conclusion. For instance, Smith's (1994: 9; cf. 13) moral realism requires the objectivity of moral judgement, where objectivity is construed as 'the idea that moral questions have correct answers, that the correct answers are made correct by objective moral facts, that moral facts are determined by circumstances, and that, by engaging in moral argument, we can discover what these objective moral facts are'. There's a lot of philosophy packed into this statement, but it looks as though Smith is committed to the thought, contra our relativist, that moral argument, or at least moral argument of the right sort, can settle moral disagreements. Indeed, for Smith (1994: 6), the notion of objectivity 'signifies the possibility of a convergence in moral views', so the prospects for his version of moral realism depend on the argument from disagreement not going through.[23] But can realists like Smith bank on the argument's failure?

Realists may argue that, in contrast to the impression one gets from the anthropological literature, there already exists substantial moral convergence. But while moral realists have often taken pretty optimistic positions on the extent of actual moral agreement (e.g. Sturgeon 1988: 229; M. Smith 1994: 188), there is no denying that there is an abundance of persistent moral disagreement. That is, on many moral issues—think of abortion and capital punishment—there is a striking failure of convergence even after protracted argument. The relativist has a ready explanation for this phenomenon: moral judgement is not objective in Smith's sense, and moral argument cannot be expected to accomplish what Smith and

[22] This way of putting the argument is at once uncontentious and contentious. It is uncontentious because it does not entail a radical methodological relativism of the sort, say, that insists there is nothing to choose between consulting an astrologer and the method of reflective equilibrium as an approach to moral inquiry (see Brandt 1959: 274–5). But precisely because of this, the empirical conjecture that moral judgements will not converge is highly contentious, since a background of methodological agreement would appear to make it more likely that moral argument could end in substantive moral agreement.

[23] Strictly speaking, a relativist need not be a 'non-factualist' about morality, since, for example, she can take it to be a moral fact that it is right for Hopi children to engage in their fatal play with small animals, and also take it to be a moral fact that it is wrong for American white children to do so. But the factualist–relativist will probably want to reject Smith's (1994: 13) characterization of moral facts as 'facts about the reasons that we all share'.

other realists think it can.[24] Conversely, the realist's task is to *explain away* failures of convergence; she must provide an explanation of the phenomena consistent with it being the case that moral judgement is objective and moral argument is rationally resolvable. For our purposes, what needs to be emphasized is that the relative merits of these competing explanations cannot be fairly determined without close discussion of actual cases. Indeed, as acute commentators with both realist (Sturgeon 1988: 230) and anti-realist (Loeb 1998: 284) sympathies have noted, the argument from disagreement cannot be evaluated by a priori philosophical means alone; what's needed, as Loeb observes, is 'a great deal of further empirical research into the circumstances and beliefs of various cultures'.

Brandt (1959: 101–2) lamented that the anthropological literature of his day did not always provide as much information on the exact contours and origins of moral attitudes and beliefs as philosophers wondering about the prospects for convergence might like. However, social psychology and cognitive science have recently produced research which promises to further discussion; the closing decades of the twentieth century witnessed an explosion of 'cultural psychology' investigating the cognitive and emotional processes of different cultures (Shweder and Bourne 1982; Markus and Kitayama 1991; Ellsworth 1994; Nisbett and Cohen 1996; Nisbett 1998; Kitayama and Markus 1999). A representative finding is that East Asians are more sensitive than Westerners to the field or context as opposed to the object or actor in their explanations of physical and social phenomena, a difference that may be reflected in their habits of ethical judgement. Here we will focus on some cultural differences found rather closer to home, differences discovered by Nisbett and his colleagues while investigating regional patterns of violence in the American North and South. We argue that these findings support Brandt's pessimistic conclusions regarding the possibility of convergence in moral judgement.

The Nisbett group's research can be seen as applying the tools of cognitive social psychology to the 'culture of honour', a phenomenon that anthropologists have documented in a variety of groups around the world. Although such peoples differ in many respects, they manifest important commonalties:

> A key aspect of the culture of honor is the importance placed on the insult and the necessity to respond to it. An insult implies that the target is weak enough to be bullied. Since a reputation for strength is of the essence in the culture of honor, the individual who insults someone must be forced to retract; if the instigator refuses, he must be punished—with violence or even death. (Nisbett and Cohen 1996: 5)

According to Nisbett and Cohen (1996: 5–9), an important factor in the genesis of southern honour culture was the presence of a herding economy. Apparently, honour cultures are particularly likely to develop where resources are liable to theft, and

[24] See Williams (1985: 136): 'In a scientific inquiry there should ideally be convergence on an answer, where the best explanation of the convergence involves the idea that the answer represents how things are; in the area of the ethical, at least at high level of generality, there is no such coherent hope.'

where the state's coercive apparatus cannot be relied upon to prevent or punish thievery. These conditions often occur in relatively remote areas where herding is the main viable form of agriculture; the 'portability' of herd animals makes them prone to theft. In areas where farming rather than herding is the principal form of subsistence, cooperation among neighbours is more important, stronger government infrastructures are more common, and resources—like decidedly unportable farmland—are harder to steal. In such agrarian social economies, cultures of honour tend not to develop. The American South was originally settled primarily by peoples from remote areas of Britain. Since their homelands were generally unsuitable for farming, these peoples have historically been herders; when they emigrated from Britain to the South, they initially sought out remote regions suitable for herding, and in such regions, the culture of honour flourished.

In the contemporary South police and other government services are widely available and herding has all but disappeared as a way of life, but certain sorts of violence continue to be more common than they are in the North. Nisbett and Cohen (1996) maintain that patterns of violence in the South, as well as attitudes towards violence, insults, and affronts to honour, are best explained by the hypothesis that a culture of honour persists among contemporary white non-Hispanic southerners. In support of this hypothesis, they offer a compelling array of evidence, including:

- demographic data indicating that (1) among southern whites, homicides rates are higher in regions more suited to herding than agriculture, and (2) white males in the South are much more likely than white males in other regions to be involved in homicides resulting from arguments, although they are *not* more likely to be involved in homicides that occur in the course of a robbery or other felony (Nisbett and Cohen 1996, ch. 2);
- survey data indicating that white southerners are more likely than northerners to believe that violence would be 'extremely justified' in response to a variety of affronts, and that if a man failed to respond violently, he was 'not much of a man' (Nisbett and Cohen 1996, ch. 3);
- legal scholarship indicating that southern states 'give citizens more freedom to use violence in defending themselves, their homes, and their property' than do northern states (Nisbett and Cohen 1996: 63).

Two experimental studies—one in the field, the other in the laboratory—are especially striking.

In the field study (Nisbett and Cohen 1996: 73–5), letters of inquiry were sent to hundreds of employers around the United States. The letters purported to be from a hard-working 27-year-old Michigan man who had a single blemish on his otherwise solid record. In one version, the 'applicant' revealed that he had been convicted for manslaughter. The applicant explained that he had been in a fight with a man who confronted him in a bar and told onlookers that 'he and my fiancée were sleeping together. He laughed at me to my face and asked me to step outside if I was man enough.' According to the letter, the applicant's nemesis was killed in the

ensuing fray. In the other version of the letter, the applicant revealed that he had been convicted of motor vehicle theft, perpetrated at a time when he needed money for his family. Nisbett and his colleagues assessed 112 letters of response, and found that southern employers were significantly more likely to be cooperative and sympathetic in response to the manslaughter letter than were northern employers, while no regional differences were found in responses to the theft letter. One southern employer responded to the manslaughter letter as follows (Nisbett and Cohen 1996: 75):

> As for your problems of the past, anyone could probably be in the situation you were in. It was just an unfortunate incident that shouldn't be held against you. Your honesty shows that you are sincere. . . . I wish you the best of luck for your future. You have a positive attitude and a willingness to work. These are qualities that businesses look for in employees. Once you are settled, if you are near here, please stop in and see us.

No letters from northern employers were comparably sympathetic.

In the laboratory study (Nisbett and Cohen 1996: 45–8) subjects—white males from both northern and southern states attending the University of Michigan—were told that saliva samples would be collected to measure blood sugar as they performed various tasks. After an initial sample was collected, the unsuspecting subject walked down a narrow corridor where an experimental confederate was pretending to work on some filing. Feigning annoyance at the interruption, the confederate bumped the subject and called him an 'asshole'. A few minutes after the incident, saliva samples were collected and analysed to determine the level of cortisol—a hormone associated with high levels of stress, anxiety and arousal, and testosterone—a hormone associated with aggression and dominance behaviour. As Figure 5.1 indicates, southern subjects showed dramatic increases in cortisol and testosterone levels, while northerners exhibited much smaller changes.

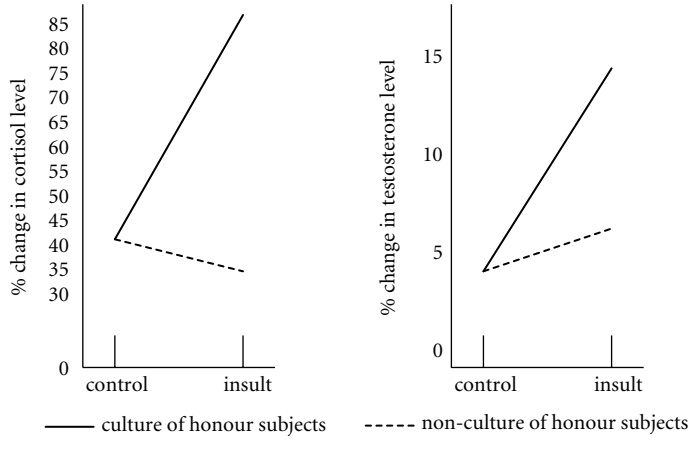

Fig. 5.1. The results of an experiment by Nisbett and Cohen in which levels of cortisol and testosterone increased much more substantially in culture of honour subjects who were insulted by a confederate

The two studies just described suggest that southerners respond more strongly to insult than northerners, and take a more sympathetic view of others who do so, manifesting just the sort of attitudes that are supposed to typify honour cultures. We think that the data assembled by Nisbett and his colleagues make a persuasive case that a culture of honour persists in the American South. Apparently, this culture affects people's judgements, attitudes, emotions, behaviour, and even their physiological responses. Additionally, there is evidence that child-rearing practices play a significant role in passing the culture of honour on from one generation to the next, and also that relatively permissive laws regarding gun-ownership, self-defence, and corporal punishment in the schools both reflect and reinforce southern honour culture (Nisbett and Cohen 1996: 60–3, 67–9). In short, it seems to us that the culture of honour is deeply entrenched in contemporary southern culture, despite the fact that many of the material and economic conditions giving rise to it no longer widely obtain.[25]

We believe that the North–South cultural differences adduced by Nisbett and colleagues support Brandt's conclusion that moral attitudes will often fail to converge, even under ideal conditions. The data should be especially troubling for the realist, for despite the differences that we have been recounting, contemporary northern and southern Americans might be expected to have rather more in common—from circumstance to language to belief to ideology—than do, say, Yanomamö and Parisians. So if there is little ground for expecting convergence under ideal conditions in the case at hand, there is probably little ground in a good many others. To develop our argument a bit further, let us revisit the idealization conditions mentioned at the beginning of this section: impartiality, full factual information, and normality.

Impartiality. One strategy favoured by moral realists concerned to explain away moral disagreement is to say that such disagreement stems from the distorting effects of individual interest (see Sturgeon 1988: 229–30); perhaps persistent disagreement doesn't so much betray deep features of moral argument and judgement as it does the doggedness with which individuals pursue their perceived advantage. For instance, seemingly moral disputes over the distribution of wealth may be due to perceptions—perhaps mostly inchoate—of individual and class interests rather than to principled disagreement about justice; persisting moral disagreement in such circumstances fails the impartiality condition, and is therefore untroubling to the moral realist.

But it is rather implausible to suggest that North–South disagreements over when violence is justified will fail the impartiality condition. There is no reason to

[25] The last clause is important, since realists (e.g. Brink 1989: 200) sometimes argue that apparent moral disagreement may result from cultures applying similar moral values to different economic conditions (e.g. differences in attitudes towards the sick and elderly between poor and rich cultures). But this explanation seems of dubious relevance to the described differences between contemporary northerners and southerners, who are plausibly interpreted as applying different values to similar economic conditions.

think that southerners would be unwilling to universalize their judgements across relevantly similar individuals in relevantly similar circumstances, as indeed Nisbett and Cohen's 'letter study' suggests. One can advocate a violent honour code without going in for special pleading.[26] We do not intend to denigrate southern values; our point is that while there may be good reasons for criticizing the honour-bound southerner, it is not obvious that the reason can be failure of impartiality, if impartiality is (roughly) to be understood along the lines of a willingness to universalize one's moral judgements.

Full and vivid awareness of relevant non-moral facts. Moral realists have argued that moral disagreements very often derive from disagreement about non-moral issues. According to Boyd (1988: 213; cf. Brink 1989: 202–3; Sturgeon 1988: 229), 'careful philosophical examination will reveal . . . that agreement on nonmoral issues would eliminate *almost all* disagreement about the sorts of moral issues which arise in ordinary moral practice'. Is this a plausible conjecture for the data we have just considered? We find it hard to imagine what agreement on non-moral facts could do the trick, for we can readily imagine that northerners and southerners might be in full agreement on the relevant non-moral facts in the cases described. Members of both groups would presumably agree that the job applicant was cuckolded, for example, or that calling someone an 'asshole' is an insult. We think it much more plausible to suppose that the disagreement resides in differing and deeply entrenched evaluative attitudes regarding appropriate responses to cuckolding, challenge, and insult.

Savvy philosophical readers will be quick to observe that terms like 'challenge' and 'insult' look like 'thick' ethical terms, where the evaluative and descriptive are commingled (see Williams 1985: 128–30); therefore, it is very difficult to say what the extent of the factual disagreement is. But this is of little help for the expedient under consideration, since the disagreement-in-non-moral-fact response apparently *requires* that one *can* disentangle factual and moral disagreement.

It is of course possible that full and vivid awareness of the non-moral facts might motivate the sort of change in southern attitudes envisaged by the (at least the northern) moral realist; were southerners to become vividly aware that their culture of honour was implicated in violence, they might be moved to change their moral outlook. (We take this way of putting the example to be the most natural one, but nothing philosophical turns on it. If you like, substitute the possibility of bloody-minded northerners endorsing honour values after exposure to the facts.) On the other hand, southerners might insist that the values of honour should be nurtured even at the cost of promoting violence; the motto 'Death before dishonour', after all, has a long and honourable history. The burden of argument, we think, lies with the

[26] The legal scholarship that Nisbett and Cohen (1996: 57–78) review makes it clear that southern legislatures are often willing to enact laws reflecting the culture of honour view of the circumstances under which violence is justified, which suggests there is at least some support among southerners for the idea that honour values should be universalizable.

realist who asserts—culture and history notwithstanding—that southerners would change their mind if vividly aware of the pertinent facts.

Freedom from abnormality. Realists may contend that much moral disagreement may result from failures of rationality on the part of discussants (Brink 1989: 199–200). Obviously, disagreement stemming from cognitive impairments is no embarrassment for moral realism; at the limit, that a disagreement persists when some or all disputing parties are quite insane shows nothing deep about morality. But it doesn't seem plausible that southerners' more lenient attitudes towards certain forms of violence are readily attributed to widespread cognitive disability. Of course, this is an empirical issue, and we don't know of any evidence suggesting that southerners suffer some cognitive impairment that prevents them from understanding demographic and attitudinal factors in the genesis of violence, or any other matter of fact. What is needed to press home a charge of irrationality is evidence of cognitive impairment independent of the attitudinal differences, and further evidence that this impairment is implicated in adherence to the disputed values in the face of the (putatively) undisputed non-moral facts. In this instance, as in many others, we have difficulty seeing how charges of abnormality or irrationality can be made without one side begging the question against the other.

We are inclined to think that Nisbett and colleagues' work represents a potent counter-example to any theory maintaining that rational argument tends to convergence on important moral issues; the evidence suggests that the North–South differences in attitudes towards violence and honour might well persist even under the sort of ideal conditions we have considered. Admittedly, our conclusions must be tentative. On the philosophical side, we have not considered every plausible strategy for 'explaining away' moral disagreement and grounding expectations of convergence.[27] On the empirical side, we have reported on but a few studies, and those we do consider here, like any empirical work, might be criticized on either conceptual or methodological grounds.[28] Finally, we should make clear what we are *not* claiming: we do not take our conclusions here—even if fairly earned—to be a 'refutation' of moral realism, in as much as there may be versions of moral realism that do not require convergence. Rather, we hope to have given an idea of the empirical work philosophers must encounter if they are to make defensible conjectures regarding moral disagreement. Our theme recurs: Responsible treatment of the empirical issues requires reference to empirical science, whatever the science is ultimately taken to show.

[27] In addition to the expedients we have considered, realists may plausibly appeal to, *inter alia*, requirements for internal coherence and the different 'levels' of moral thought (theoretical versus popular, abstract versus concrete, general versus particular) at which moral disagreement may or may not be manifested. Brink (1989: 197–210) and Loeb (1998) offer valuable discussions with considerably more detail than we offer here, Brink manifesting realist sympathies and Loeb tending towards anti-realism.

[28] We think Nisbett and Cohen will fare pretty well under such scrutiny. See Tetlock's (1999) favourable review.

5. Thought Experiments

Ethical reflection is often held to involve comparing general principles and responses to particular cases; commitment to a principle may compel the renunciation of a particular response, or commitment to a particular response may compel modification or renunciation of a general principle (Brandt 1959: 244–52; Rawls 1971: 20–1, 49). This emphasis on particular cases is not peculiar to ethics: 'intuition pumps' or 'thought experiments' have long been central elements of philosophical method (Dennett 1984: 17–18). In the instances we consider here, a thought experiment presents an example, typically a hypothetical example, in order to elicit some philosophically telling response; if a thought experiment is successful, it may be concluded that competing theories must account for the resulting response.[29] To extend the imagery of experimentation, responses to thought experiments are supposed to serve an evidential role in philosophical theory choice; the responses are data competing theories must accommodate.[30]

In ethics, one—we do not say the only—familiar rendering of the methodology is this: if an audience's ethical responses to a thought experiment can be expected to conflict with the response a theory prescribes for the case, the theory has suffered a counter-example. For instance, it is often claimed that utilitarian prescriptions for particular cases will conflict with the ethical responses many people have to those cases (e.g. Williams 1973: 99). The ethics literature is rife with claims to the effect that 'many of us' or 'we' would respond in a specified way to a given example, and such claims are often supposed to have philosophical teeth.[31] But who is this 'we'? And how do philosophers know what this 'we' thinks?

Initially, it doesn't look like 'we' should be interpreted as 'we philosophers'. The difficulty is not that this approach threatens a sampling error, although it is certainly true that philosophers form a small and peculiar group. Rather, the problem is that philosophers can be expected to respond to thought experiments in ways that reflect their theoretical predilections: utilitarians' responses to a thought

[29] There are substantive questions as to what sorts of responses to thought experiments may properly constrain philosophical theory choice. For example, what level of reflection or cognitive elaboration is required: are the responses of interest 'pre-theoretical intuitions' or 'considered judgements'? We will have something to say about this, but in terminology we will mostly favour the generic 'responses', which we mean to be neutral regarding issues such as cognitive elaboration.

[30] This analogy with science is not unique to our exposition. Singer (1974: 517; cf. 493) understands Rawls's (1971) method of reflective equilibrium as 'leading us to think of our particular moral judgments as data against which moral theories are to be tested'. As Singer (1974: 493 ff.) notes, in earlier treatments Rawls (1951) made the analogy with scientific theory choice explicit. We needn't hazard an interpretation of Rawls, but only observe that our analogy is not philosophically eccentric.

[31] For appeals of this kind, see Blum (1994: 179); G. Strawson (1986: 87–9); P. Strawson (1982: 68); Wallace (1994: 81–2); Williams (1973: 99–100; 1981: 22).

experiments might be expected to plump for maximizing welfare, integrity and loyalty be damned, while the responses of Aristotelians and Kantians might plump in the opposite direction. If so, the thought experiment can hardly be expected to *resolve* the debate, since philosophers' responses to the example are likely to *reflect* their position in the debate.

The audience of appeal often seems to be some variant of 'ordinary folk' (see Jackson 1998: 118, 129; Jackson and Pettit 1995: 22–9; Lewis 1989: 126–9). Of course, the relevant folk must possess such cognitive attainments as are required to understand the case at issue; very young children are probably not an ideal audience for thought experiments. Some philosophers may want to insist that the relevant responses are the 'considered judgements' or 'reflective intuitions' of people with the training required to see 'what is philosophically at stake'. But there is peril in insisting that the relevant cognitive attainments be some sort of 'philosophical sophistication'. Once again, if the responses are to help adjudicate between competing theories, the responders must be more or less theoretically neutral, but this sort of neutrality, we suspect, is rather likely to be vitiated by philosophical education. (Incredibly enough, informal surveys suggest that *our* students are overwhelmingly ethical naturalists!)

However exactly the philosophically relevant audience is specified, there are empirical questions that must be addressed in determining the philosophical potency of a thought experiment. In science, not all experiments produce data of evidentiary value; sampling errors and the failure of experimental designs to effectively isolate variables are two familiar ways in which experiments go wrong. Data resulting from such experiments is tainted, or without evidential value; analogously, in evaluating responses to a thought experiment, one needs to consider the possibility of taint. In particular, when deciding what philosophical weight to give a response to a thought experiment, philosophers need to determine the origins of the response. What features of the example are implicated in a response—are people responding to the substance of the case, or the style of exposition? What features of the audience are implicated in a response—do different demographic groups respond to an example differently? Such questions raise the following concern: ethical responses to thought experiments may be strongly influenced by ethically irrelevant characteristics of example and audience. Whether a characteristic is ethically relevant is a matter for philosophical discussion, but determining the status of a particular thought experiment also requires empirical investigation of its causally relevant characteristics; responsible philosophical discussion cannot rely on guesswork in this regard. We shall now give two examples illustrating our concerns about tainted origins, one corresponding to each of the two questions just asked.

Tversky and Kahneman presented subjects with the following problem:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been

> proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die. (Tversky and Kahneman 1981: 453)

On the first version of the problem most subjects thought that Program A should be adopted. But on the second version most chose Program D, despite the fact that the outcome described in A is identical to the one described in C. The disconcerting implication of this study is that ethical responses may be strongly influenced by the manner in which cases are described or framed. Many effects of framing differences, such as that between 200 of 600 people being saved and 400 of 600 dying, we are strongly inclined to think, are ethically irrelevant influences on ethical responses (compare Horowitz 1998; Sinnott-Armstrong 2005). Unless this sort of possibility can be confidently eliminated, one should hesitate to rely on responses to a thought experiment for adjudicating theoretical controversies. Again, such possibilities can only be eliminated through systematic empirical work.[32]

Audience characteristics may also affect the outcome of thought experiments. Haidt and associates (1993: 613) presented stories about 'harmless yet offensive violations of strong social norms' to men and women of high and low socio-economic status (SES) in Philadelphia (USA), Porto Alegre, and Recife (both in Brazil). For example: 'A man goes to the supermarket once a week and buys a dead chicken. But before cooking the chicken, he has sexual intercourse with it. Then he cooks it and eats it' (Haidt *et al.* 1993: 617). Lower SES subjects tended to 'moralize' harmless and offensive behaviours like that in the chicken story: these subjects were more inclined than their privileged counterparts to say that the actor should be 'stopped or punished', and more inclined to deny that such behaviours would be 'OK' if customary in a given country (Haidt *et al.* 1993: 618–19). The point is not that lower SES subjects are mistaken in their moralization of such behaviours while the urbanity of higher SES subjects represents the most rationally defensible response. To recall our previous discussion of moral disagreement, the difficulty is deciding which of the conflicting responses to privilege, when both sorts of responses may be the function of more or less arbitrary cultural factors.

[32] Some authors—most notably Baron (1994)—have argued that the distorting influences of 'heuristics and biases' like those uncovered in the recent psychological literature on reasoning, judgement, and decision-making are widespread in everyday ethical reflection. For overviews of the relevant psychological literature, see Nisbett and Ross (1980); Kahneman *et al.* (1982); Baron (2001).

In presenting the Haidt group's work to philosophical audiences, our impression is that they typically decline to moralize the offensive behaviours, and we ourselves share their tolerant attitude. But of course philosophical audiences—by virtue of educational attainments if not stock portfolios—are overwhelmingly high SES. Haidt's work suggests that it is a mistake for a philosopher to say, as Jackson (1998: 32 ff.; cf. 37) does, that 'my intuitions reveal the folk conception in as much as I am reasonably entitled, as I usually am, to regard myself as typical'. The question is: Typical of what demographic? Are philosophers' ethical responses to thought experiments determined by the philosophical substance of the examples, or by cultural idiosyncrasies that are very plausibly thought to be ethically irrelevant? Once again, until such possibilities are ruled out by systematic empirical investigation, the philosophical heft of a thought experiment is open to question.[33]

The studies just described raise provocative questions about *how* responses to thought experiments are generated, but there may be equally provocative questions about *what* responses people actually have. And, to sound our now familiar theme, this question is one not credibly answered by guesswork. Indeed, we suspect that philosophical speculations about what responses to thought experiments are conventional may be wrong surprisingly often. We'll now report on one study conducive to such suspicions.

One of the most famous of philosophical conundrums, that of determinism and responsibility, can be derived—on one way of formulating the difficulty—from the juxtaposition of three claims that are individually quite plausible, but seem impossible to hold jointly:

(MRT) *Moral responsibility thesis*: Human beings are sometimes morally responsible for their behaviour.

(CT) *Causal thesis*: All human behaviour is linked to antecedent events by deterministic causal laws. (See Scanlon 1988: 152.)

(PAP) *Principle of alternate possibilities*: A 'person is morally responsible for what he has done only if he could have done otherwise'. (See Frankfurt 1988: 1.)

Here's one way of putting it: If CT is true, it looks as though it is never the case that people could have done otherwise, but then, given PAP, MRT must be false.[34] There

[33] We applaud Jackson's (1998: 36–7) advocacy of 'doing serious public opinion polls on people's responses to various cases'. However, we expect this may be necessary more often than Jackson imagines. According to Jackson (1998: 37), 'Everyone who presents the Gettier cases [which are well-known epistemology thought experiments] to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases.' Yet Weinberg *et al.* (2002) found that responses to epistemology thought experiments like the Gettier cases varied with culture and SES; this suggests that philosophers need to be more systematic in their fieldwork.

[34] Our formulation is meant to be quite standard. Kane (2002*a*: 10) observes that statements of the difficulty in terms of alternative possibilities have dominated modern discussion. A recently prominent formulation proceeds not in terms of PAP, but by way of an 'ultimacy condition', which holds that an actor is responsible for her behaviour only if she is its 'ultimate source' (see McKenna 2001, esp.

are three standard responses to this trilemma. Two sorts of incompatibilists hold that MRT and CT cannot be held simultaneously: hard determinists (see Smart 1961: 303–6) reject MRT,[35] while libertarians (e.g. Kane 1996) insist that CT admits of exceptions in the case of human behaviour, and are thus able to maintain MRT. Compatibilists, on the other hand, assert that MRT and CT can be simultaneously maintained; one well-known expedient is to reject PAP, and insist that people may be legitimately held responsible even when they could not have done otherwise (see Frankfurt 1988: 1–12).

The literature is voluminous, and the proffered solutions range from controversial to deeply unsatisfying; indeed, there is heated disagreement as to what exactly the problem is (Dennett 1984: 1–19). Discretion being the best part of valour, we won't review the arguments here. Given our present concerns, we instead consider objections to the effect that compatibilism is in some sense badly counter-intuitive. One way of forming this complaint is to say that people's 'reactive attitudes'—ethical responses like anger, resentment, guilt, approbation, admiration, and the like—manifest a commitment to incompatibilism.[36] Here is Galen Strawson (1986: 88) on what he calls the 'incompatibilist intuition':

> The fact that the incompatibilist intuition has such power for us is as much a natural fact about cogitative beings like ourselves as is the fact of our quite unreflective commitment to the reactive attitudes. What is more, the roots of the incompatibilist intuition lie deep in the . . . reactive attitudes. . . . The reactive attitudes enshrine the incompatibilist intuition.[37]

Let's do a little unpacking. On Strawson's (1986: 31; cf. 2, 84–8) rendering, incompatibilism is the view that the falsity of determinism is a necessary condition for moral responsibility. To suggest that the 'incompatibilist intuition' is widespread, then, may be thought to imply that people's (possibly tacit) body of moral beliefs includes commitment to the claim that CT is incompatible with MRT.[38]

40–1). This does not impact the present discussion, however. First, notice that although some may maintain an ultimacy requirement and reject PAP, the two commitments need not be incompatible; Kane (1996, 2002*b*) holds them both. Secondly, as should become clear, the empirical work we describe below is relevant to both formulations.

[35] As Kane (2002*a*: 27–32) observes, relatively few philosophers have been unqualifiedly committed to hard determinism; Smart's (1961) views on responsibility, for example, are complex.

[36] Peter Strawson (1982) did the pioneering philosophical work on the reactive attitudes; he appears to reject the suggestion that such attitudes manifest a commitment to something in the spirit of incompatibilism.

[37] G. Strawson puts the point rather emphatically, but similar observations are commonplace. Cf. Nagel (1986: 113, 125); Kane (1996: 83–5).

[38] There is again a question about the scope of 'people'; Strawson's reference to 'natural facts' may suggest that he is making a boldly pancultural attribution, but he might be more modestly attributing the theory only to those people who embody something like the 'Western ethical tradition'. We will not attempt to decide the interpretative question, because the empirical work we describe troubles even the more modest claim.

This is an empirical claim. Moreover, it is an empirical claim that looks to entail predictions about people's moral responses. What are the responses in question?

Like many other philosophers making empirical claims about human cognition and behaviour, Strawson says relatively little about what predictions he thinks his claims entail. We won't put predictions in Strawson's mouth; instead, we'll consider one prediction that looks to follow from positing an incompatibilist intuition, at least on the familiar rendering of incompatibilism we've followed. Attributing a widespread commitment to an incompatibilist intuition is plausibly thought to involve the following prediction: for cases where the actor is judged unable to have done otherwise, people will not hold the actor responsible for what she has done.[39] In as much as this prediction is a good one, people should respond to thought experiments depicting an actor unable to do otherwise by abjuring attributions of responsibility and the associated reactive attitudes.

In a compatibilist spirit inspired by the work of Harry Frankfurt (1988), Woolfolk, Doris, and Darley (forthcoming) hypothesized that observers may hold actors responsible even when the observers judge that the actors could not have done otherwise, at least in cases where the actors appear to manifest 'identification'. Very roughly, the idea is that the actor is identified with a behaviour—and is therefore responsible for it—to the extent she 'embraces' the behaviour (or its motive), or performs it 'wholeheartedly'.[40] Woolfolk *et al.*'s suspicion was, in effect, that people's (possibly tacit) theory of responsibility is, contra Galen Strawson and others, compatibilist.

In one of the Woolfolk *et al.* studies, subjects read a story about two married couples vacationing together. According to the story, one of the vacationers has discovered that his wife is having an affair with his opposite number in the foursome; on the flight home, the vacationers' plane is hijacked, and armed hijackers order the cuckold to shoot the man who has been having an affair with his wife. In a 'low identification' variation, the story contained the following material:

> Bill was horrified. At that moment Bill was certain about his feelings. He did *not* want to kill Frank, even though Frank was his wife's lover. But although he was appalled by the situation and beside himself with distress, he reluctantly placed the pistol at Frank's temple and proceeded to blow his friend's brains out.

Conversely, in a 'high identification' variation, the embittered cuckold embraces his opportunity:

> Despite the desperate circumstances, Bill understood the situation. He had been presented with the opportunity to kill his wife's lover and get away with it. And at that moment Bill

[39] G. Strawson (1986: 25–31; 2002) may favour formulations in terms of ultimacy rather than PAP (see n. 33 above). This doesn't affect our argument, since the empirical work we recount below looks to trouble a prediction formulated in terms of ultimacy as well as the alternative possibilities formulation we favour.

[40] For some discussion, see Velleman (1992); Bratman (1996); Watson (1996); Doris (2002: 140–6).

was certain about his feelings. He wanted to kill Frank. Feeling no reluctance, he placed the pistol at Frank's temple and proceeded to blow his friend's brains out.

Consistent with Woolfolk and colleagues' hypothesis, the high-identification actor was judged more responsible, more appropriately blamed, and more properly subject to guilt than the low-identification actor.[41]

It is tempting to conclude that at least for the Woolfolk group's subjects (philosophy and psychology undergraduates at the University of California and Rutgers University), the incompatibilist intuition does not appear to be deeply entrenched. But at this point the incompatibilist will be quick to object: the above study may suggest that responsibility attributions are influenced by identification, but it says nothing about commitment to the incompatibilist intuition, because subjects may not have believed that the actor could not have done otherwise, and subjects therefore cannot be interpreted as attributing responsibility in violation of PAP. People may think that even when coerced, actors 'always have a choice'; in the classic 'your money or your life' scenario, the person faced with this unpleasant dilemma can always opt for her life. (We hasten to remind anyone tempted in such a bull-headed direction that the disjunct need not be exclusive!)

To address this objection, Woolfolk *et al.* attempted to elevate perceived constraint to the 'could not have done otherwise' threshold:

> The leader of the kidnappers injected Bill's arm with a 'compliance drug'—a designer drug similar to sodium pentathol, 'truth serum.' This drug makes individuals unable to resist the demands of powerful authorities. Its effects are similar to the impact of expertly administered hypnosis; it results in total compliance. To test the effects of the drug, the leader of the kidnappers shouted at Bill to slap himself. To his amazement, Bill observed his own right hand administering an open-handed blow to his own left cheek, although he had no sense of having willed his hand to move. The leader then handed Bill a pistol with one bullet in it. Bill was ordered to shoot Frank in the head. . . . when Bill's hand and arm moved again, placing the pistol at his friend's temple, Bill had no feeling that he had moved his arm to point the gun; it felt as though the gun had moved itself into position. Bill thought he noticed his finger moving on the trigger, but could not feel any sensations of movement. While he was observing these events, feeling like a puppet, passively observing his body moving in space, his hand closed on the pistol, discharging it and blowing Frank's brains out.

Strikingly, subjects appeared willing to attribute responsibility to the shooter even here: once again, a high-identification actor was judged more responsible, more appropriately blamed, and more properly subject to guilt than a low-identification actor. No doubt this is not the most 'naturalistic' scenario, but neither is it outlandish by philosophical standards. And it certainly looks to be a case where the actor would be perceived to fail the standard for responsibility set by PAP.[42] Indeed,

[41] Woolfolk *et al.* (forthcoming) obtained similar results for the prosocial behaviour of kidney donation: an identified actor was credited for making a donation even when heavily constrained.

[42] It also looks as though the actor fails an ultimacy condition (see nn. 34 and 39 above).

Woolfolk *et al.* found that subjects were markedly less likely to agree to statements asserting that the actor 'was free to behave other than he did', and 'could have behaved differently than he did', than they were in the case of simple coercion described above. These results look to caution against positing a widespread commitment to the incompatibilist intuition. Deciding empirical issues concerning habits of responsibility attribution will not, of course, decide the philosophical dispute between compatibilists and incompatibilists. Yet in so far as the incompatibilist is making claims to the effect that compatibilists cannot accommodate entrenched habits of moral response, the empirical evidence is entirely relevant.

Once more, some philosophers may insist that the responses of interest are not the relatively unschooled or intuitive responses of experimental subjects like the Woolfolk group's undergraduates, but the tutored judgements of philosophers. We've already given some reasons for regarding this strategy with suspicion, but it seems to us especially problematic for the particular case of responsibility. Philosophical arguments about responsibility, it seems to us, often lean heavily on speculation about everyday practice. For example, Peter Strawson's (1982: 64, 68) extremely influential exposition repeatedly stresses the importance of reactive attitudes in 'ordinary inter-personal relationships'. While it may not be too much of a stretch to imagine that philosophers sometimes indulge in such relationships, it *is* a stretch to suppose that they are the only folk who do so. It is very plausible to argue—as indeed those who have deployed something like the incompatibilist intuition have done—that the contours of the everyday practice of responsibility attribution serve as a (defeasible) constraint on philosophical theories of responsibility: if the theory cannot accommodate the practice, it owes, at a bare minimum, a debunking account of the practice. One might insist that philosophical theorizing about responsibility is not accountable to ordinary practice, but this is to make a substantial break with important elements of the tradition.

There are a couple of ways in which philosophers can avoid the sorts of empirical difficulties we have been considering. First, they can deny that responses to particular cases have evidential weight in ethical theory choice, as some utilitarians—unsurprisingly given the rather startling implications of their position—have been inclined to do (e.g. Kagan 1989: 10–15; Singer 2000, p. xviii). Alternatively, they can appeal to the results of thought experiments in an expository rather than an evidential role; for example, a thought experiment might be used by an author to elucidate her line of reasoning without appealing to the responses of an imagined audience like 'many of us'. To some philosophers, such solutions will seem rather methodologically draconian, threatening to isolate ethical theory from the experience of ethical life (see Williams 1985: 93–119, esp. 116–19). But our point here is less grand: many users of thought experiments in ethics apparently have been—and we strongly suspect will continue to be—in the business of forwarding an imagined consensus on their thought experiments as evidence in theory choice. For these philosophers we offer the following methodological prescription: a credible philosophical

methodology of *thought* experiments must be supplemented by a cognitive science of thought experiments that involves systematic investigation with *actual* experiments. There are just too many unanswered questions regarding the responses people have, and the processes by which they come to have them. We've no stake in any particular answers to such questions. What we do have a stake in, as we have throughout, is the observation that responsible answers to such questions will be informed by systematic empirical investigation.

6. Conclusion

We needn't linger on goodbyes; the main contours of our exposition should by now be tolerably clear. We have surveyed four central topics in ethical theory where empirical claims are prominent: character, moral motivation, moral disagreement, and thought experiments. We have argued that consideration of work in the biological, behavioural, and social sciences promises substantive philosophical contributions to controversy surrounding such topics as virtue ethics, internalism, moral realism, and moral responsibility. If our arguments are successful, we have also erected a general methodological standard: philosophical ethics can, and indeed must, interface with the human sciences.[43]

References

Annas, J. (1993). *The Morality of Happiness*. New York: Oxford University Press.

Anscombe, G. E. M. (1958). 'Modern Moral Philosophy'. *Philosophy*, 33: 1–19.

Aristotle (1984). *The Complete Works of Aristotle*, ed. J. Barnes. Princeton: Princeton University Press.

Athanassoulis, N. (2000). 'A Response to Harman: Virtue Ethics and Character Traits'. *Proceedings of the Aristotelian Society*, 100: 215–22.

Audi, R. (1995). 'Acting from Virtue'. *Mind*, 104: 449–71.

Baron, J. (1994). 'Nonconsequentialist Decisions'. *Behavioral and Brain Sciences*, 17: 1–42.

[43] For much valuable feedback, we are grateful to audiences at the Moral Psychology Symposium at the 2001 Society for Philosophy and Psychology meetings, the Empirical Perspectives on Ethics Symposium at the 2001 American Philosophical Association Pacific Division meetings, and a series of lectures on philosophy and cognitive science held at the Australian National University in July 2002—especially Louise Antony, Daniel Cohen, Frank Jackson, Michael Smith, and Valerie Tiberius. Thanks to Daniel Guevara, Jerry Neu, Alva Noë, and especially Don Loeb, Shaun Nichols, and Adina Roskies for comments on earlier drafts.

——— (2001). *Thinking and Deciding*, 3rd edn. Cambridge: Cambridge University Press.
Bechara, A., Damasio, H., and Damasio, A. R. (2000). 'Emotion, Decision Making and the Orbitofrontal Cortex'. *Cerebral Cortex*, 10: 295–307.
Becker, L. C. (1998). *A New Stoicism*. Princeton: Princeton University Press.
Bennett, W. J. (1993). *The Book of Virtues: A Treasury of Great Moral Stories*. New York: Simon & Schuster.
Blackburn, S. (1998). *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press.
Blair, R. J. (1995). 'A Cognitive Developmental Approach to Morality: Investigating the Psychopath'. *Cognition*, 57: 1–29.
——— Jones, L., Clark, F., and Smith, M. (1997). 'The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?' *Psychophysiology*, 34: 192–8.
Blum, L. A. (1994). *Moral Perception and Particularity*. Cambridge: Cambridge University Press.
Bok, H. (1996). 'Acting without Choosing'. *Noûs*, 30: 174–96.
Boyd, R. N. (1988). 'How to Be a Moral Realist', in G. Sayre-McCord (ed.), *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.
Brandt, R. B. (1954). *Hopi Ethics: A Theoretical Analysis*. Chicago: University of Chicago Press.
——— (1959). *Ethical Theory: The Problems of Normative and Critical Ethics*. Englewood Cliff, NJ: Prentice-Hall.
——— (1970). 'Traits of Character: A Conceptual Analysis'. *American Philosophical Quarterly*, 7: 23–37.
Bratman, M. E. (1996). 'Identification, Decision, and Treating as a Reason'. *Philosophical Topics*, 24: 1–18.
Brink, D. O. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.
Campbell, J. (1999). 'Can Philosophical Accounts of Altruism Accommodate Experimental Data on Helping Behaviour?' *Australasian Journal of Philosophy*, 77: 26–45.
Cooper, J. M. (1999). *Reason and Emotion: Essays on Ancient Moral Psychology and Ethical Theory*. Princeton: Princeton University Press.
Damasio, A. R., Tranel, D., and Damasio, H. (1990). 'Individuals with Sociopathic Behavior Caused by Frontal Damage Fail to Respond Autonomically to Social Stimuli'. *Behavioral Brain Research*, 41: 81–94.
Daniels, N. (1979). 'Wide Reflective Equilibrium and Theory Acceptance in Ethics'. *Journal of Philosophy*, 76: 256–84.
Darley, J. M., and Batson, C. D. (1973). 'From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior'. *Journal of Personality and Social Psychology*, 27: 100–8.
D'Arms, J., and Jacobson, D. (2000). 'Sentiment and Value'. *Ethics*, 110: 722–48.
Darwall, S. L. (1983). *Impartial Reason*. Ithaca, NY: Cornell University Press.
——— (1989). 'Moore to Stevenson', in Robert Cavalier, James Gouinlock, and James Sterba (eds.), *Ethics in the History of Philosophy*. London: Macmillan.
——— Gibbard, A., and Railton, P. (eds.) (1997). *Moral Discourse and Practice: Some Philosophical Approaches*. New York: Oxford University Press.
Deigh, J. (1999). 'Ethics', in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*. Cambridge: Cambridge University Press.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, Mass.: MIT Press.

Dent, N. J. H. (1975). 'Virtues and Actions'. *Philosophical Quarterly*, 25: 318–35.

DePaul, M. (1999). 'Character Traits, Virtues, and Vices: Are There None?', in *Proceedings of the 20th World Congress of Philosophy*, i. Bowling Green, Ohio: Philosophy Documentation Center.

Doris, J. M. (1996). 'People Like Us: Morality, Psychology, and the Fragmentation of Character'. Ph.D. diss., University of Michigan, Ann Arbor.

—— (1998). 'Persons, Situations, and Virtue Ethics'. *Noûs*, 32: 504–30.

—— (2002). *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

—— and Stich, S. P. (2001). 'Ethics', in *The Encyclopedia of Cognitive Science*, philosophy ed. D. Chalmers. London: Macmillan Reference.

Ellsworth, P. C. (1994). 'Sense, Culture, and Sensibility', in H. Markus and S. Kitayama (eds.), *Emotion and Culture: Empirical Studies in Mutual Influence*. Washington: American Psychological Association.

Firth, R. (1952). 'Ethical Absolutism and the Ideal Observer Theory'. *Philosophy and Phenomenological Research*, 12: 317–45.

Flanagan, O. (1991). *Varieties of Moral Personality: Ethics and Psychological Realism*. Cambridge, Mass.: Harvard University Press.

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Frankena, W. K. (1976). 'Obligation and Motivation in Recent Moral Philosophy', in K. E. Goodpaster (ed.), *Perspectives on Morality: Essays of William K. Frankena*. Notre Dame, Ind.: University of Notre Dame Press.

Frankfurt, Harry (1988). *The Importance of What we Care About: Philosophical Essays*. Cambridge: Cambridge University Press.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, Mass.: Harvard University Press.

Gilbert, D. T., and Malone, P. S. (1995). 'The Correspondence Bias'. *Psychological Bulletin*, 117: 21–38.

Goldman, A. I. (1993). 'Ethics and Cognitive Science'. *Ethics*, 103: 337–60.

Haidt, J., Koller, S., and Dias, M. (1993). 'Affect, Culture, and Morality; or, Is it Wrong to Eat your Dog?'. *Journal of Personality and Social Psychology*, 65: 613–28.

Haney, C., Banks, W., and Zimbardo, P. (1973). 'Interpersonal Dynamics of a Simulated Prison'. *International Journal of Criminology and Penology*, 1: 69–97.

Hare, R. D. (1993). *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York: Pocket Books.

Hare, R. M. (1952). *The Language of Morals*. Oxford: Oxford University Press.

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

—— (1999). 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error'. *Proceedings of the Aristotelian Society*, 99: 315–31.

—— (2000). 'The Nonexistence of Character Traits'. *Proceedings of the Aristotelian Society*, 100: 223–6.

Hart, D., and Killen, M. (1999). 'Introduction: Perspectives on Morality in Everyday Life', in M. Killen and D. Hart (eds.), *Morality in Everyday Life: Developmental Perspectives*, paperback edn. Cambridge: Cambridge University Press.

Hill, T. E. (1991). *Autonomy and Self-Respect*. Cambridge: Cambridge University Press.
Horowitz, T. (1998). 'Philosophical Intuitions and Psychological Theory', in M. DePaul and W. Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. Lanham, Md.: Rowman & Littlefield.
Hume, D. (1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edn. Oxford: Oxford University Press.
—— (1978). *A Treatise of Human Nature*, 2nd edn. Oxford: Oxford University Press.
Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.
Hutcheson, F. (1738). *An Enquiry into the Original of our Ideas of Beauty and Virtue, in Two Treatises*. London: D. Midwinter.
Irwin, T. H. (1988). 'Disunity in the Aristotelian Virtues'. *Oxford Studies in Ancient Philosophy*, supp. vol., 61–78.
Isen, A. M., and Levin, P. F. (1972). 'Effect of Feeling Good on Helping: Cookies and Kindness'. *Journal of Personality and Social Psychology*, 21: 384–8.
Jackson, F. (1994). 'Armchair Metaphysics', in J. O'Leary Hawthorne and M. Michael (eds.), *Philosophy in Mind*. Dordrecht: Kluwer.
—— (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. New York: Oxford University Press.
—— and Pettit, P. (1995). 'Moral Functionalism and Moral Motivation'. *Philosophical Quarterly*, 45: 20–40.
Johnson, M. (1993). *Moral Imagination: Implications of Cognitive Science for Ethics*. Chicago: University of Chicago Press.
Jones, E. E. (1990). *Interpersonal Perception*. New York: W. H. Freeman.
Kagan, S. (1989). *The Limits of Morality*. Oxford: Oxford University Press.
Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.
—— (2002*a*). 'Introduction: The Contours of Contemporary Free Will Debates', in R. Kane (ed.), *The Oxford Handbook of Free Will*. New York: Oxford University Press.
—— (2002*b*). 'Some Neglected Pathways in the Free Will Labyrinth', in R. Kane (ed.), *The Oxford Handbook of Free Will*. New York: Oxford University Press.
Keneally, T. (1982). *Schindler's List*. New York: Simon & Schuster.
Kim, J. (1988). 'What Is "Naturalized Epistemology"?', in J. Tomberlin (ed.), *Philosophical Perspectives, ii*: *Epistemology*. Atascadero, Calif.: Ridgeway.
Kitayama, S., and Markus, H. R. (1999). '*Yin* and *Yang* of the Japanese Self: The Cultural Psychology of Personality Coherence', in D. Cervone and Y. Shoda (eds.), *The Coherence of Personality: Social–Cognitive Bases of Consistency, Variability, and Organization*. New York: Guilford Press.
Kupperman, J. J. (2001). 'The Indispensability of Character'. *Philosophy*, 76: 239–50.
Larmore, C. E. (1987). *Patterns of Moral Complexity*. Cambridge: Cambridge University Press.
Leming, J. S. (1997*a*). 'Research and Practice in Character Education: A Historical Perspective', in A. Molnar (ed.), *The Construction of Children's Character: Ninety-Sixth Yearbook of the National Society for the Study of Education*, p. II. Chicago: University of Chicago Press.
—— (1997*b*). 'Whither Goes Character Education? Objectives, Pedagogy, and Research in Character Education Programs'. *Journal of Education*, 179: 11–34.

Lewis, D. (1970). 'How to Define Theoretical Terms'. *Journal of Philosophy*, 67: 427–46.
—— (1972). 'Psychophysical and Theoretical Identifications'. *Australasian Journal of Philosophy*, 50: 249–58.
—— (1989). 'Dispositional Theories of Value'. *Proceedings of the Aristotelian Society*, suppl. vol., 63: 113–37.
Loeb, D. (1998). 'Moral Realism and the Argument from Disagreement'. *Philosophical Studies*, 90: 281–303.
Louden, R. B. (1984). 'On Some Vices of Virtue Ethics'. *American Philosophical Quarterly*, 21: 227–36.
McDowell, J. (1978). 'Are Moral Requirements Hypothetical Imperatives?' *Proceedings of the Aristotelian Society*, suppl. vol., 52: 13–29.
—— (1979). 'Virtue and Reason'. *The Monist*, 62: 331–50.
—— (1987). *Projection and Truth in Ethics*. Lindley Lecture. Kansas: University of Kansas.
MacIntyre, A. (1984). *After Virtue*, 2nd edn. Notre Dame, Ind.: University of Notre Dame Press.
McKenna, M. (2001). 'Source Incompatibilism, Ultimacy, and the Transfer of Non-Responsibility'. *American Philosophical Quarterly*, 38: 37–51.
Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.
Margolis, E., and Laurence, S. (1999). *Concepts*. Cambridge, Mass.: MIT Press.
Markus, H. R., and Kitayama, S. (1991). 'Culture and the Self: Implications for Cognition, Emotion, and Motivation'. *Psychological Review*, 98: 224–53.
Mathews, K. E., and Cannon, L. K. (1975). 'Environmental Noise Level as a Determinant of Helping Behavior'. *Journal of Personality and Social Psychology*, 32: 571–7.
Merritt, M. (1999). 'Virtue Ethics and the Social Psychology of Character'. Ph.D. diss., University of California, Berkeley.
—— (2000). 'Virtue Ethics and Situationist Personality Psychology'. *Ethical Theory and Moral Practice*, 3: 365–83.
Milgram, S. (1974). *Obedience to Authority*. New York: Harper & Row.
Mischel, W. (1968). *Personality and Assessment*. New York: Wiley.
Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.
Nagel, T. (1986). *The View from Nowhere*. New York: Oxford University Press.
Nichols, S. (2002). 'How Psychopaths Threaten Moral Rationalism; or, Is it Irrational to Be Amoral?' *The Monist*, 85: 285–304.
—— (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.
Nisbett, R. E. (1998). 'Essence and Accident', in J. M. Darley and J. Cooper (eds.), *Attribution and Social Interaction: The Legacy of Edward E. Jones*. Washington: American Psychological Association.
—— and Cohen, D. (1996). *Culture of Honor: The Psychology of Violence in the South*. Boulder, Colo.: Westview Press.
—— and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.
Nucci, L. (1986). 'Children's Conceptions of Morality, Social Conventions and Religious Prescription', in C. Harding (ed.), *Moral Dilemmas: Philosophical and Psychological Reconsiderations of the Development of Moral Reasoning*. Chicago: Precedent Press.
Nussbaum, M. C. (1999). 'Virtue Ethics: A Misleading Category?' *Journal of Ethics*, 3: 163–201.
Peterson, D. R. (1968). *The Clinical Study of Social Behavior*. New York: Appleton-Century-Crofts.

Quine, W. v. O. (1969). 'Epistemology Naturalized', in Quine, *Ontological Relativity and Other Essays*. New York: Columbia University Press.
Railton, P. (1986*a*). 'Facts and Values'. *Philosophical Topics*, 14: 5–31.
—— (1986*b*). 'Moral Realism'. *Philosophical Review*, 95: 163–207.
—— (1989). 'Naturalism and Prescriptivity'. *Social Philosophy and Policy*, 7: 151–74.
—— (1995). 'Made in the Shade: Moral Compatibilism and the Aims of Moral Theory'. *Canadian Journal of Philosophy*, suppl. vol., 21: 79–106.
Rawls, J. (1951). 'Outline of a Decision Procedure for Ethics'. *Philosophical Review*, 60: 167–97.
—— (1971). *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.
Rosati, C. S. (2000). 'Brandt's Notion of Therapeutic Agency'. *Ethics*, 110: 780–811.
Roskies, A. (2003). 'Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy" '. *Philosophical Psychology*, 16: 51–66.
Ross, L., and Nisbett, R. E. (1991). *The Person and the Situation: Perspectives of Social Psychology*. Philadelphia: Temple University Press.
Saver, J. L., and Damasio, A. R. (1991). 'Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy Due to Ventromedial Frontal Damage'. *Neuropsychologia*, 29: 1241–9.
Scanlon, T. M. (1988). 'The Significance of Choice', in S. M. McMurrin (ed.), *The Tanner Lectures on Human Values*, viii. Salt Lake City: University of Utah Press.
Sher, G. (1998). 'Ethics, Character, and Action', in E. F. Paul, F. D. Miller, and J. Paul (eds.), *Virtue and Vice*. Cambridge: Cambridge University Press.
Sherman, N. (1989). *The Fabric of Character: Aristotle's Theory of Virtue*. New York: Oxford University Press.
Shweder, R. A., and Bourne, E. J. (1982). 'Does the Concept of the Person Vary Cross-Culturally?', in A. J. Marsella and G. M. White (eds.), *Cultural Conceptions of Mental Health and Therapy*. Boston, Mass.: Reidel.
Singer, P. (1974). 'Sidgwick and Reflective Equilibrium'. *The Monist*, 58: 490–517.
—— (2000). *Writings on an Ethical Life*. New York: HarperCollins.
Sinnott-Armstrong, W. P. (2005). 'Moral Intuitionism Meets Empirical Psychology', in T. Horgan and M. Timmons (eds.), *Metaethics After Moore*. New York: Oxford University Press.
Smart, J. J. C. (1961). 'Free-Will, Praise and Blame'. *Mind*, 70: 291–306.
Smith, Adam. (2002). *The Theory of Moral Sentiments*. New York: Cambridge University Press.
Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.
Stevenson, C. L. (1944). *Ethics and Language*. New Haven: Yale University Press.
—— (1963). *Facts and Values*. New Haven: Yale University Press.
Stich, S. (1993*a*). 'Naturalizing Epistemology: Quine, Simon and the Prospects for Pragmatism', in C. Hookway and D. Peterson (eds.), *Philosophy and Cognitive Science*, Royal Institute of Philosophy, suppl. 34. Cambridge: Cambridge University Press.
—— (1993*b*). 'Moral Philosophy and Mental Representation', in M. Hechter, L. Nadel, and R. E. Michod (eds.), *The Origin of Values*. New York: de Gruyter.
Strawson, G. (1986). *Freedom and Belief*. Oxford: Oxford University Press.
—— (2002). 'The Bounds of Freedom', in R. Kane (ed.), *The Oxford Handbook of Free Will*. New York: Oxford University Press.
Strawson, P. (1982). 'Freedom and Resentment', in G. Watson (ed.), *Free Will*. New York: Oxford University Press.

Sturgeon, N. L. (1988). 'Moral Explanations', in G. Sayre-McCord (ed.), *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.
Sumner, W. G. (1934). *Folkways*. Boston: Ginn.
Svavarsdóttir, S. (1999). 'Moral Cognitivism and Motivation'. *Philosophical Review*, 108: 161–219.
Tetlock, P. E. (1999). 'Review of *Culture of Honor: The Psychology of Violence in the South*'. *Political Psychology*, 20: 211–13.
Turiel, E., Killen, M., and Helwig, C. (1987). 'Morality: Its Structure, Functions, and Vagaries', in J. Kagan and S. Lamb (eds.), *The Emergence of Morality in Young Children*. Chicago: University of Chicago Press.
Tversky, A., and Kahneman, D. (1981). 'The Framing of Decisions and the Psychology of Choice'. *Science*, 211: 453–63.
Velleman, J. D. (1992). 'What Happens When Someone Acts?' *Mind*, 101: 461–81.
Vernon, P. E. (1964). *Personality Assessment: A Critical Survey*. New York: Wiley.
Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, Mass.: Harvard University Press.
Watson, G. (1990). 'On the Primacy of Character', in Owen Flanagan and Amélie Oksenberg Rorty (eds.), *Identity, Character, and Morality: Essays in Moral Psychology*. Cambridge, Mass.: MIT Press.
—— (1996). 'Two Faces of Responsibility'. *Philosophical Topics*, 24: 227–48.
Weinberg, J., Nichols, S., and Stich, S. (2002). 'Normativity and Epistemic Intuitions'. *Philosophical Topics*, 29: 429–60.
Westermarck, E. (1906). *Origin and Development of the Moral Ideas*, 2 vols. New York: Macmillan.
Williams, B. A. O. (1973). 'A Critique of Utilitarianism', in J. J. C. Smart and B. A. O. Williams, *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.
—— (1981). *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.
—— (1985). *Ethics and the Limits of Philosophy*. Cambridge, Mass.: Harvard University Press.
—— (1993). *Shame and Necessity*. Berkeley: University of California Press.
Woods, M. (1986). 'Intuition and Perception in Aristotle's Ethics'. *Oxford Studies in Ancient Philosophy*, 4: 145–66.
Woolfolk, R. L., Doris, J. M. (2002). 'Rationing Mental Health Care: Parity, Disparity, and Justice'. *Bioethics*, 16: 469–85.
—— —— and Darley, J. M. (forthcoming). 'Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility.' *Cognition*.

CHAPTER 28

PHILOSOPHY OF BIOLOGY

PHILIP KITCHER

1. Origins and Evolution

In the middle decades of the twentieth century, biology was hardly visible in twentieth-century philosophy of science. Apart from occasional references to vitalism, a few discussions of teleology, and J. H. Woodger's valiant (but scholastic) attempt to make evolutionary biology fit within the frame provided by standard logical empiricism (Woodger 1937), philosophers concentrated their attentions on physics and psychology, with chemistry, anthropology, and history receiving less discussion, but still substantially more air time than the life (and earth) sciences. Ironically, at the same time, biological science was undergoing major transformations, first in the modern synthesis (the integration of Darwinian evolutionary theory and classical genetics, accomplished in the 1930s and 1940s through the work of R. A. Fisher, Sewall Wright, J. B. S. Haldane, Theodosius Dobzhansky, G. G. Simpson, and Ernst Mayr), and later in the post-war birth of molecular biology. These advances, particularly the latter, made biology the glory science of the second half of the century—at most major universities today there are far more biology students than specialists in other sciences, and, at many, enrolments in biology exceed all other sciences combined. Sooner or later, philosophers of science were bound to notice.

In the 1960s there were early pioneers, Morton Beckner (1959), Marjorie Grene (1959), and T. A. Goudge (1961), but those who established the philosophy of biology as a thriving, independent subdiscipline of philosophy of science came a generation

later. In the writings of David Hull, Michael Ruse, and William Wimsatt, philosophers encountered wide-ranging discussions of a variety of problems, informed by technical details of biology at a level that had long been attained in the parallel philosophical study of the physical sciences. By and large, however, the principal philosophical focus was the first major transition in mid-century biology, the forging of the evolutionary synthesis. Molecular biology, despite its obvious increasing hegemony in the life sciences, was relatively neglected.

Philosophers approaching biological materials naturally brought with them tools and concepts that had been fashioned in general studies of science, studies nourished by favourite examples from physics. It was not hard to see that the tools and concepts did not always fit the biological cases, and that casual claims made about science in general often could not be sustained. Hence the turn to biology frequently brought lessons for the general philosophy of science. Sometimes, indeed, the presentation of biological complexities undermined popular claims in metaphysics or in moral philosophy. Most philosophers of biology, however, were not content to regard their field as a new laboratory in which parts of philosophy could be tested. Like their colleagues in the philosophy of physics, they were excited by the opportunity to engage in theoretical disputes within biology—they wrote for biologists as well as for philosophers. From the mid-1970s to the present, there have been serious collaborations between biologists and philosophers, and, in general, practising biologists have tended to see the work of philosophers of biology as relevant to their concerns and to appraise it more favourably than their physicist colleagues welcome work in philosophy of physics. Moreover, because late twentieth-century biology has often been applied to issues of great social concern—as in debates about the genetics of intelligence (and, more broadly, in controversies in behavioural genetics), in human sociobiology, in evolutionary psychology, and the Human Genome Project—philosophers have had the opportunity to help clarify, and even resolve, questions of obvious practical significance.

Philosophy of biology has thus evolved in a number of directions, sometimes attempting to illuminate the general philosophy of science, sometimes offering new perspectives on old philosophical problems, sometimes entering the theoretical fray within biology, sometimes participating in controversies about the social implications of biological findings. Although it's useful to consider these four varieties of philosophical work, it would be artificial and misguided to attempt to classify every article or book as exemplifying just one of them. Authors writing on the character of species, for example, may have an interest both in sorting out a biological controversy and in enriching philosophical discussion about natural kinds. Sometimes, as we'll see below, the attempt to engage a socially vexed question prompts a new theoretical debate that demands philosophical analysis. The evolution of the philosophy of biology has revealed the interaction of different pressures, in many combinations. What follows is a review of what has interested philosophers who have usually started with a focus on evolutionary biology, and

have been variously prodded by a philosophical thesis here, a novel piece of biology there, and an arousing of public interest somewhere else.[1]

1.1 The Status (and Structure?) of Evolutionary Theory

One of the obvious reasons for concentrating on Darwin's theory of evolution by natural selection is that it doesn't look much like those respectable theories within physical science that have attracted philosophical attempts at reconstruction. According to logical empiricism, scientific theories are axiomatic systems, some of whose axioms employ special vocabulary ('theoretical terms'), and the best theories are those that generate a broad class of consequences that admit of test ('observation sentences'). Informal attempts to cast Darwin's achievement in this mould tend to view his theory as consisting in a 'principle of natural selection', as if this were the sole axiom of evolutionary theory. Efforts to articulate this principle then proceed to formulate it as the claim

(1) Heritable traits that increase the fitness of their bearers increase in frequency in a population

or something similar. At this point, a decision must be made about the concept of fitness. Philosophers have pursued two main options. The first supposes that "fitness" is a theoretical term, in the classical logical empiricist sense, and that its meaning is specified by correspondence rules (M. Williams, 1970; Rosenberg 1982, 1983); but this faces the obvious difficulty that biological practice rarely, if ever, provides general principles about fitness (even about fitness in types of environments), offering instead a bundle of specific claims about the relative fitnesses of highly specific traits in carefully characterized situations. The second approach attempts a general definition of fitness. Here, an obvious way to identify fitness is to appeal to the measures that field biologists actually employ, and to equate relative fitness with relative number of progeny (and here there are options depending on whether one counts the descendants in the first, or some later, generation). That strategy, however, invites the retort that Darwin's theory has now become a triviality (or, as creationists love to proclaim, a 'tautology')—that what it says is

(2) Heritable traits that increase the number of descendants left by their bearers increase in frequency in a population.

In the 1970s some philosophers attempted to evade this discouraging conclusion by offering a slightly different account of fitness: they took relative fitnesses to be not

[1] I should note explicitly that this review is idiosyncratic; I have concentrated on those aspects of contemporary philosophy of biology that have struck me as most interesting. I suspect that others would draw the map rather differently.

actual relative numbers of progeny, but *expected* relative numbers of progeny (Mills and Beatty 1979; Brandon 1990). Thus (1) was transformed into

(3) Heritable traits that are expected to increase the number of descendants left by their bearers increase in frequency in a population.

Darwin's theory now stands as the barely empirical claim that expectation values are actualized.

Partly under pressure from creationist challenges, philosophers have done better, and they have done so by breaking free of the idea that all areas of flourishing science should be reconstructed by viewing them as offering theories in the classical logical empiricist sense. If one wants to insist that the only viable notion of a theory is that of an axiomatic system (with the familiar empiricist conditions about theoretical and observational vocabulary), then the point is that one can have major scientific achievements that don't provide 'theories'. A more common view has been that the logical empiricist conception of theory is too narrow. From the late 1970s on, a number of commentators urged the merits of using the semantic conception of theories (according to which a theory is identified with a family of models) as a means of reconstructing Darwin (Beatty 1980*b*; Thompson 1983; Lloyd 1983, 1988). Others remained closer to the framework of logical empiricism by urging that Darwin offered a pattern of explanation (Kitcher 1982*b*, 1985*a*) or a general causal mechanism (Sober 1984) that was used again and again in the explanation of biological phenomena. Any of these approaches was able to provide a more satisfying account of what Darwin's classic work (1859) accomplishes. The immense detail of the chapters on comparative anatomy and morphology, on embryology, the fossil record, and perhaps most of all on biogeography, serves as the illustration of how Darwin's general framework of viewing all organisms as related by descent with modification and his invocation of natural selection as an agent of transspecific modification can be used to explain phenomena that would otherwise be puzzling. One can develop this approach in terms of the fecundity of an abstract model type, in terms of the unifying power of a few explanatory patterns, or by recognizing the omnipresence of some general causal mechanisms.

So what is the structure of evolutionary theory, either as formulated initially by Darwin or as it emerges in the work of his successors? That strikes me as a bad question, one born of a general logical empiricist project that we ought to abandon. Scientists practise their craft by doing a wide variety of things, observing, experimenting, refining techniques, offering predictions, intervening, giving explanations, sometimes combining apparently disparate phenomena under a single perspective. Philosophers who want to 'reconstruct' this diverse practice can usually do so in a variety of ways, and different modes of reconstruction can be valuable for distinct purposes. The much-maligned logical empiricist conception of theories as axiomatic systems is sometimes useful precisely because axiomatizations lend themselves to discussion about issues of independence of assumptions (to cite one obvious example).

I suggest that there is *no* genuine question about *the* structure of scientific theories. Philosophers have a bundle of techniques that are more or less useful in answering particular purposes. The failure of the logical empiricist conception of theory in the Darwinian case lay in the facts that (*a*) there was no serious question to which it usefully lent itself, and (*b*) it was a disaster for clarifying the epistemological status of Darwinian claims (especially in a context in which religious fundamentalists delighted in making that status as obscure as possible).

2. The Units of Selection Debate

One of the first areas in which the philosophy of biology engaged with a debate in theoretical biology was the controversy over the units of selection. The sources of the dispute lay in discussions of the early twentieth century. Darwin's theory seemed to imply that organisms would never have a heritable tendency to behave in ways that promote the reproductive success of other members of their populations at reproductive cost to themselves. Yet behavioural biologists seemed to find recurrent instances of just this type of 'altruistic' behaviour.[2] Hence arose the task of explaining the possibility of altruism in a Darwinian world. An initially attractive solution proposed that the types of behaviour in question endured because of a benefit to the group to which the actor belongs (perhaps conceived as a local population, or as the entire species). During the 1960s that suggestion came under intense scrutiny: a number of writers showed that appeals to group benefits were much more problematic than had been appreciated and that the puzzling phenomena could be understood by taking selection to operate on individuals—or even on genes (Hamilton 1964; Maynard Smith 1964; G. Williams 1966).

The last thought was developed with great rhetorical flair by Richard Dawkins in his widely read book *The Selfish Gene* (1976). There Dawkins summarized the theoretical advances of a fertile decade and recast them as demonstrating that selection really acts on genes. By this he intended not merely that one can keep track of evolutionary changes by recording the frequencies of gene variants (alleles) across the generations, but that the process of selection should be seen as one in which alleles have advantages or disadvantages (in the pertinent environments), and that these advantages and disadvantages are the causes of subsequent differences in allelic frequencies. In Dawkins's vision, the packaging of genes on chromosomes,

[2] According to the biological definition, an organism A acts altruistically towards another organism B just in case A's action increases the reproductive success (the number of offspring reaching maturity in the next generation) of B and diminishes the reproductive success of A.

the embedding of the genetic material in cells, and the aggregation of cells into multicellular organisms should all be understood as the expression of good strategies for genes to lever themselves into future generations.

After Dawkins, biologists and philosophers could formulate the general 'units of selection problem'. The 'orthodox Darwinian' view takes organisms to be units of selection; confused pre-1960 discussions often invoked groups (superorganismic entities) as units of selection; radical Darwinians (Dawkins and George Williams) claim that genes are the units of selection; who is right? There were early criticisms of Dawkins's proposal; Stephen Jay Gould argued that genes were not 'visible' to natural selection, suggesting that only manifest traits make a difference to survival, attraction of mates, fecundity, and the other conditions on which the transmission of genes into the next generation depends (Gould 1980*a*); Robert Brandon subsequently tried to make Gould's concerns more precise, by supposing that manifest traits will typically 'screen off' underlying genetic characteristics (Brandon 1984). Biological discussion quickly attracted philosophical attention to the formulations traded by the contending participants. David Hull (1981) drew a useful distinction between *replicators* (those entities that are transmitted unchanged across generations) and *interactors* (entities that engage in causal processes affecting reproductive success); if the unit of selection is conceived as a replicator, then genes are an obvious candidate; if the unit of selection is conceived as an interactor, then Dawkins's thesis is far less immediate.

An article by Elliott Sober and Richard Lewontin (1982) and the subsequent book by Sober (1984) greatly advanced the discussion.[3] Sober and Lewontin conceded that genes can be used for 'bookkeeping' in processes of natural selection, but they denied that genes can figure as causal agents. In support of their latter claim they introduced an important example. It is possible for there to be two alleles at a locus (A and *a*) so that the homozygotes (AA, *aa*) are lethal (organisms with these combinations die young) while the heterozygotes (A*a*) thrive. Under these circumstances, both alleles will be viewed as having equal fitness (because, in each generation, there are equal frequencies of each), so that a genic view will disclose no selection. In each generation, however, selection is plainly occurring, for all the homozygotes (half the population, since matings between heterozygotes yield each of the homozygotes one-quarter of the time) die before reaching maturity. Dawkins's vision thus obliterates important causal facts about the situation.

Sober continued by developing a causal criterion that he invoked to resolve issues about the units of selection. Causes, he suggests, must raise the probability of effects, and they must do so in all causally relevant background contexts. More exactly:

(4) C causes E only if, for any causally relevant background condition B_i, $Pr(E/C \& B_i) \geq Pr(E/B_i)$, with the inequality holding strictly in at least one case.

[3] In my judgement, the importance of Sober's (1984) account of natural selection is comparable to the celebrated monographs by Hans Reichenbach that have been so seminal in philosophy of physics. As will be apparent in what follows, I don't agree with Sober's conclusions.

(4) can be deployed to sharpen the Sober–Lewontin argument, for, as the contrived example shows, the effects of alleles on reproductive success may be positive or negative depending on other genotypic features (the effect of A varies according to whether or not it is accompanied with another A or an *a*). Moreover, as Sober showed, (4) allows one to make sense of the notion of group selection. His discussion advocated a *pluralistic hierarchical* view of selection. When we consider a natural selection process, the units of selection may sometimes be genes, sometimes organisms, sometimes groups; in some complicated processes, selection can even act at several distinct levels, so that selection for genes is countered (or reinforced) by selection for organisms (or groups).[4]

Although other philosophers and biologists (Gould 1982; Lloyd 1988; Wimsatt 1981; Brandon 1990; Godfrey-Smith and Lewontin 1993) may dissent from the details of Sober's argument, preferring their own substitutes for his framework (centred on the use of (4)), the pluralistic hierarchical view of selection has become the dominant position in the field. Nevertheless, it seems to me to be incorrect. Although his original (1976) accepted the commitment to a *real unit of selection*, Dawkins's (1982) contained hints of a different view, one according to which there is no fact of the matter. Kim Sterelny and I have attempted to articulate this view (Sterelny and Kitcher 1988; see also Waters 1991). In response to Sober's critique of genetic selection, we point out that principle (4) fails to govern the most celebrated examples of natural selection, all of which average across causally relevant background contexts. The famous case of industrial melanism, for example, introduces selective pressures on moth populations that abstract from local details—for even in woods that are polluted there will be clumps of trees that are unaffected and in which the speckled (non-melanic) form has an advantage in escaping predation. Further, as we show, it is possible to attribute genic fitnesses in ways that will capture the causal facts of selection, so long as one is careful to identify the environment in the right way; thus, in Sober's example of the lethal homozygotes, part of the environment of each allele is the allele with which it is paired—an A promotes reproductive success in the context of *a*, and detracts from reproductive success in the context of another A. (The point is further articulated in the important Godfrey-Smith and Lewontin 1993, although the authors try to defend their hierarchical view by adverting to an ill-defined conception of causation as sometimes acting at different levels.) So, in contrast to the pluralist hierarchical view, Sterelny and I propose a pluralist conventionalism. Where our opponents suppose that for each selection process there is a unique correct causal description, typically invoking a single causal level (with different levels picked out in different instances) and sometimes recognizing multiple levels, we suggest that any selection process can usually be described in a number of different ways, with the genic perspective

[4] Strictly speaking, Sober's account treats selection as being 'for' *traits* of genes (or organisms, or groups), but I'll gloss over the niceties here; see Sober (1984) for details.

being most widely available, and that the choice among modes of description is purely pragmatic.

To simplify, one can formulate our position as the claim that the units of selection debate over-interprets Darwin's metaphor. Organisms are born, they mate, they reproduce, and they die.[5] Darwin offered us a way of thinking about the aggregate results of such individual occurrences, and contemporary evolutionary theory has demonstrated how to make precise mathematical models that subsume a myriad of causal details. It seems at least disputable that there should be privileged assignments of fitness values to particular entities, assignments that capture the 'causal facts' about 'levels of selection'; for when a hawk picks a moth off a tree, although we have a causal interaction between a bird and an insect, there's no place at which *selection* points; we can recognize the causal interaction without forgoing the right to describe the general process of which it is a tiny part in any mode that seems most convenient. The talk of selection is a *tool* which we introduce to sum up a complex array of causal facts, and we should fashion the tool so as best to suit our concerns.

For most participants in the debate, however, pluralistic conventionalism is seen not as a way of avoiding dubious metaphysics but as a refusal to respond to the complete causal facts. In recent years, discussions on this point seem to have stalled. The main recent contribution to the units of selection controversy has been the effort by Sober and the evolutionary theorist David Sloan Wilson to rehabilitate the notion of group selection (Sober and Wilson 1998). Sober and Wilson have used their well-articulated account of group selection to generate a synthetic approach to the issue of altruism (see Section 6 below), as well as to advance the claims of the pluralistic hierarchical view. From my perspective, they have offered biologists a richer choice of models of evolutionary processes, adding to the repertoire from which we can legitimately choose in tackling the complexities of birth, reproduction, and death.

3. Concepts and Methods of Evolutionary Theorizing

Many other issues concerning evolutionary biology have attracted philosophical attention, either because, like the units of selection controversy, they are debated by prominent biologists, or because they bear on long-standing philosophical questions. In this section I'll briefly look at four that seem particularly important.

In 1979 Stephen Jay Gould and Richard Lewontin published an influential essay suggesting that evolutionary theory was in the grip of a 'Panglossian paradigm'

[5] Of course, some organisms are asexual, and don't mate at all. For them the story is simpler.

(Gould and Lewontin 1979). They charged that evolutionary theorists were too keen to see the action of selection everywhere in nature, and in consequence that they were lulled into accepting accounts of the evolution of traits, including forms of behaviour, on the basis of inadequate evidence. Gould and Lewontin were particularly moved by examples from sociobiology (see Section 6 below), where, they suggested, a penchant for 'just so stories' often led researchers astray.

Darwin famously claimed that natural selection has been the chief but not the sole agent of modification (1859: 6). Responses to Gould and Lewontin quite naturally asked what alternative causal mechanisms these authors intended to invoke (Mayr 1983). The ensuing debate identified two main issues: First, when selection acts, can we always think of it as producing an optimal outcome? Secondly, to what extent is the operation of selection supplemented or countered by other causes? Although answers to these questions are often conflated with theses about the units of selection, it is important to recognize that they are logically independent of that debate: one can maintain any position on which entities (if any) are 'real' units of selection, while holding any view about selection and optimality or any view about modes of evolutionary causation. The urge to conflate probably arises from the fact that the most prominent advocate of genic selection, Richard Dawkins, is also one of the most outspoken defenders of the scope and power of natural selection (Dawkins 1987).

Biologists and philosophers have articulated with some care the conditions under which one can deploy optimality models in understanding the operation of natural selection. A succession of articles has clarified the kinds of constraints to which claims of optimization must be subject (Oster and Wilson 1978; Maynard Smith 1978; Beatty 1980*a*; Dupré 1987; Orzack and Sober 1994; Sober 1998). The second question has been much more controversial. Some biologists and philosophers have insisted on the power of natural selection (Dawkins 1987; Dennett 1995) and have condemned Gould, Lewontin, and their fellow-travellers for mystery-mongering. The strongest claims for alternative agents of evolutionary change have emerged from attempts to expose sources of order not recognized in contemporary Darwinism. Gould and Lewontin had already emphasized the possibility of deeply entrenched patterns of development (*Baupläne*—a concept introduced by German biologists who are often dismissed by orthodox Darwinians), and Gould's first book (his often neglected 1978) focused on historical and current explorations of that possibility. Subsequent contributions by philosophers (Wimsatt 1986; Beurton *et al.* 2000) and by biologists (Raff 1996; Meinhardt 1998) have pursued this theme, and Stuart Kauffman (1993) is a particularly thorough and well-developed attempt to investigate whether there are sources of biological order that need to be integrated with Darwinian orthodoxy.

I turn now to a second debate that is closely linked to the adaptationism controversy. One moral that might be drawn from (Gould and Lewontin 1979) is that the connections between evolutionary biology and developmental biology must be

more precisely articulated (see also Lewontin 1974*a*). To ignore the processes of development that underlie a structure or trait is to risk offering an absurd 'just so story'—as if one suggested that jutting chins have been selected for their value in male displays (overlooking the fact that the chin emerges as a by-product of growth in two developmental fields). Scholars incensed by the influence of casual adaptationist thinking on nature–nurture controversies have been particularly moved to suggest that the source of the trouble is the way in which Darwinian theory has ignored development, and they have called for a new synthesis (Ho and Saunders 1984; Ho and Fox 1988). Particularly influential has been Susan Oyama's proposal of 'developmental systems theory' (Oyama 1985), which is explicitly intended to deliver the 'stake-in-the-heart move' to recurrent claims about the limits placed on us by our biological nature.

Oyama's work has inspired biologists like Russell Gray and philosophers such as Paul Griffiths and Kim Sterelny (Griffiths and Gray 1994; Sterelny and Griffiths 1999). In the strongest versions (those of Oyama, Griffiths, and Gray) our standard gene-centred versions of evolutionary theory should give way to a new style of analysis, one that takes the developmental system as central. This new perspective will recognize that what organisms inherit are not only stable chunks of DNA, but also other important molecules (the proteins that play a crucial role in early development, for example) and enduring aspects of the environment. There are connections between the celebration of developmental systems theory and other calls for the reform of evolutionary theory (for example, the 'dialectical biology' of Levins and Lewontin 1985).

At the heart of developmental systems theory is a principle of causal democracy that biologists should be happy to accept. (Indeed one common reaction is to insist that this principle is already well established in orthodox theorizing.) That principle recognizes that the manifest traits of organisms are not products of their genes alone, but emerge from the intricate interactions among pieces of DNA, other molecules, and environmental causes at multiple levels. Developmental systems theory, however, aims to go beyond this undisputed interactionism, and the principal challenge to it consists in asking for a precise specification of just the ways in which orthodox interactionism is deficient. My own view is that the insights of the proponents of developmental systems theory can be accommodated without any serious departure from orthodoxy—other, perhaps, than a commitment to keep interactionism firmly in focus—and that we cannot hope to drive a stake in the heart of all ventures in applying biology in socially harmful ways (Kitcher 2000).

The two issues so far considered have emerged first within evolutionary theory, although resolution of them has consequences for philosophical positions (Dennett 1995). The third topic of this section is a more purely philosophical problem. Biology is full of attributions of functions to traits, structures, organs, and forms of behaviour. Perhaps before the nineteenth century those functional claims could be understood in terms of the design of the creator who intended that plants

and animals should be able to satisfy their needs, but, in a post-Darwinian world, they are much harder to understand. This was already appreciated before the philosophy of biology came of age, and C. G. Hempel and Ernest Nagel both devoted some attention to analysing the character of functional explanation (Hempel 1965; Nagel 1979). The task of clarifying what functional claims mean has continued to exercise philosophers.

The two most important rival proposals were both advanced in the 1970s. Robert Cummins (1973) suggested that to give a functional analysis of an item (trait, organ, structure) is to provide a causal decomposition of the production of that item. By contrast, Larry Wright (1973) argued that claims of the form 'The function of X is Y' should be understood as meaning that Y is both an effect of X and also the explanation of why X is there. Both views encounter apparent difficulties. Cummins's account seems to allow for the attribution of function to items that play a causal role in virtually any kind of process—thus we can use his analysis to identify the function that an outcrop of rock plays in the flow of water down a mountainside. Similarly, Wright's proposal licenses some peculiar functional claims; in a telling example offered by Christopher Boorse (1976) if a leaky hose in a laboratory causes a scientist to collapse from asphyxiation before he can fix it, then we're committed to the odd idea that the function of the leak is to asphyxiate scientists.

Although both suggestions were subsequently developed in the philosophical literature, Wright's received the major share of attention and his 'aetiological approach' has become the orthodox treatment of functional and teleological notions. An important elaboration was offered by Ruth Millikan, who presented a complex account of 'proper functions' explicitly linking the notion of function to evolution (Millikan 1984). Thus, for Millikan and her successors (Neander 1991; Godfrey-Smith 1994; Sober 1984; Mitchell 1995; Bigelow and Pargetter 1987), the functions of X are those effects of X whose past instances have played a causal role in the evolution of X under natural selection. Using the hoary example of the heart, we may say that the function of hearts is to pump blood, meaning thereby that pumping blood is one of the things that hearts do and that the ability of past hearts to do that was advantageous to organisms that had rudimentary hearts and thus played a role in the process of natural selection through which hearts evolved.

Although the aetiological account, thus articulated, seems to accord with many features of biological practice (see e.g. Gould and Vrba 1982) it is at odds with a celebrated distinction drawn by the ethologist Niko Tinbergen, who explicitly separated questions about evolutionary origin from questions about function (Tinbergen 1963). Since Millikan's detailed treatment, philosophers have considered the temporal location of the selection process through which the attribution of function is supposed to be grounded. At one extreme, one can demand that the function of X is Y only if Y played a role in the selection process through which X originated; at the other, one can look into the future, and suppose that Y explains the presence of X in the generations that will succeed the present (Bigelow and

Pargetter 1987; for criticism, see Mitchell 1993). Godfrey-Smith has offered a middle view that has the advantage of allowing us to draw Tinbergen's distinction; his 'modern history' theory of functions proposes that the appropriate selective regime be one in the recent past that has maintained X into the present (Godfrey-Smith 1994). Even with this clarification in place, however, the aetiological view still faces questions about the extent to which selection played a role in the maintenance of the appropriate item (Kitcher 1993).

There is a broader concern. Functional attributions abound in areas of biology where evolution is far from the dominant focus, and in which researchers would cheerfully confess their ignorance of evolutionary detail. In molecular studies in physiology, for example, there are routine claims about the function of various enzymes, despite the fact that the evolutionary pressures on the pertinent features of past organisms are swathed in obscurity. Precisely in these areas, the Wright–Millikan approach appears forced, and Cummins's proposal—which would allow researchers to deploy the kinds of causal considerations they typically advance to support their functional ascriptions—seems to do better. Yet, as seen briefly above, Cummins's analysis is too liberal. I've suggested that we can modify that analysis and integrate it with the aetiological view by supposing that functional analysis only proceeds against the background of a general view of selective pressures, whether or not we know how to articulate a history of selection in the case at hand (Kitcher 1993). Godfrey-Smith has countered that a unified vision of functions is not so easily achieved (Godfrey-Smith 1993).

The fourth and last topic from evolutionary theory that I want to discuss here connects both with biological debates and with very broad questions in metaphysics. In 1942 Ernst Mayr remedied an important deficiency in Darwinian evolutionary theory by offering an account of the notion of species (that Darwin had taken for granted). According to Mayr's 'biological species concept', species taxa are clusters of populations each of which would freely interbreed with others (if present in nature at the same place at the same time) and would not freely interbreed with other populations that fall outside the cluster (Mayr 1942, 1963). Although Mayr was successful in convincing most of his biological colleagues to adopt his definition, it has been the object of much discussion from the 1940s to the present. One obvious issue concerns the counterfactual condition: there are instances in which it simply is impossible to bring candidate populations into geographical contact with one another (think, for example, of extinct organisms), and in such cases Mayr's concept is hard to apply with conviction. Another centres on the existence of species that reproduce asexually, with respect to which Mayr has often advised that biologists deploy morphological criteria that have been found to coincide with species divisions among the closest sexually reproducing relatives of the asexual organisms in question.

The biological species concept was embedded within a more general approach to classifying organisms—'evolutionary systematics'—that drew higher distinctions (into genera, classes, families, and so forth) partly on the basis of morphological

similarity, partly on the basis of evolutionary relationships. For over four decades, systematists—those who study biological classification—have debated the merits of this approach, and it has come to be seen as a compromise between two polar positions. One of these, 'numerical taxonomy', proposed to classify organisms by using a large number of characteristics for which numerical measures could be assigned and looking for clusters in the high-dimensional space defined by those characteristics. The other, 'cladistics', resolutely insists on evolutionary relationships, even if the resulting classifications turn out to be at odds with morphological similarities or past biological practice. Cladists have developed various precise compendia of rules for identifying how evolutionary kinship is to be assessed, and there are currently a number of different versions of the position.

During the 1970s an evolutionary biologist, Michael Ghiselin, and a philosopher, David Hull, mounted a serious challenge to the biological species concept, one that had ramifications for the more general issues about classification (Ghiselin 1974; Hull 1976). Ghiselin and Hull contended that species are not 'classes' but 'individuals'. As Hull articulated the point in a seminal essay (Hull 1978), we should think of species taxa as individuated by the role they play within the genealogy of living things; they are segments of the tree of life, bound together by actual relations of reproduction and descent. A consequence that Hull explicitly noted is that species, once extinct, cannot recur. The Ghiselin–Hull proposal captured biological attention because of its kinship with cladistic approaches to systematics (although many cladists have developed different concepts of species) and because it seemed to allow for higher-order processes of selection of the sort proposed in articulating an 'expansion' of evolutionary theory (Eldredge 1985; Gould 1980*b*, 2002). It also provoked considerable philosophical discussion.

One uncontroversial moral for philosophy has been that much of the literature on natural kinds, even some that is most influential, has been ill-designed to cope with the complexities of biological examples (see Dupré 1981, 1993). Beyond this, there has been a wide-ranging debate about the thesis that species are individuals, about how to individuate species taxa, and about the implications of a view of species for our general understanding of evolutionary theory. Although the Ghiselin–Hull thesis has been widely accepted, it seems, at first sight, to rest on a confusion: for it contrasts an ontological possibility (species are individuals rather than collections, or sets) with a semantic possibility (species names can be defined without reference to spatio-temporal markers). I have suggested that the confusion is real, and that the important point is to identify species taxa in ways that make essential reference to their historical origins (Kitcher 1984*a*, 1989). Framing the issue in this way enables us to make sense of the wide variety of proposals for individuating species (see the essays in Ereshevsky 1992). One option would then be to decide that one of the contending proposals is correct. From my perspective, that would be a mistake, and we ought to honour different approaches to species, advanced with different biological purposes in mind; thus, contra Hull, I think there's a use for a

species concept that enables us to make sense of the recurrence of species, and that this concept would naturally be employed by biomedical researchers concerned with the possibility that *the same pathogen* might be produced again (Kitcher 1984*a*).

As I've already noted, the Ghiselin–Hull proposal was linked both to cladistic approaches to classification of organisms and to suggestions about expanding Darwinism. On the latter front, it seems that the logical connections are very loose. Whether or not we can make sense of species selection, and whether or not appeals to species selection are needed to make sense of the history of life, are matters for careful analysis of selection processes and of actual cases in palaeontology; they aren't settled by drawing semantic distinctions and making arguments in ontology. By contrast, the work of philosophers has played a valuable role in the fierce disputes among rival systematists, where Hull has served both as patient elucidator of rival views and as chronicler of the campaigns (Hull 1979, 1988). Cladistic methodology has also benefited from the careful study provided by Elliott Sober of the ways in which various desiderata are deployed in the development of a genealogical account of life (Sober 1988).

For many biologists, probably for almost all, the questions of classification and the principles that should guide it are arcane, even 'philosophical' in the pejorative sense. The discussions of species concepts from the 1970s to the present are probably more important for refining philosophical proposals in traditional areas of metaphysics than they are for reforming or aiding biology (so that Dupré 1993 may be the most enduring contribution of a large literature; see also Splitter 1988). One remarkable feature of the discussion has been the almost invariable insistence on the need to fit classificatory concepts (primarily that of species) to the needs of evolutionary theory, as if there were no other areas of biology whose projects should be considered. The protracted 'species debate' thus highlights the 'evolutionary chauvinism' that has dominated philosophy of biology. I now turn to philosophical ventures in other parts of the life sciences.

4. The Neglected Elephant

As noted in Section 1, philosophy of biology came of age a decade or so after two major transitions in the life sciences, and, as just claimed, it has focused particularly on one of these. Yet, from the 1960s to the present, an increasing proportion of biologists have been working in the areas transformed by the second—the emergence and acceleration of molecular biology—so that today's undergraduate philosophy of biology class often *introduces* most of the students within it to the *scientific* material about which it philosophizes, even though those enrolled have substantial

backgrounds in biology. What they know about, of course, is things like DNA replication and the roles of various enzymes in metabolic pathways. Molecular biology sits in the classroom like a neglected elephant.

One possible explanation for this might be that not all parts of the sciences are philosophically interesting. Thus one might claim that evolutionary theory has attracted so much attention because it raises large philosophical issues, that it is a sad fact that evolution occupies less biological attention than it used, and that philosophers should be impervious to this change of fashion. I'll try to argue that the premiss is false, that molecular biology poses interesting philosophical questions, but, even were that not so, philosophers might have important work to do in articulating theoretical issues that arise in molecular studies (as they have done with respect to the units of selection, the adaptationism controversy and to disputes in systematics).

The most prominent philosophical work that touches on molecular biology addresses the issue of whether the life sciences can be reduced to the physical sciences. Since the 1960s philosophers of science have debated issues about the reduction of 'higher-order' sciences to more fundamental disciplines. In pursuing these questions, they typically employed an influential model of reduction articulated by Ernest Nagel (1962): scientific theories are viewed as axiomatic systems, whose axioms consist of laws of nature, and reduction is effected by deriving the axioms of the reduced theory from the axioms of the reducing theory, possibly with the aid of 'bridge principles' that specify the referents of terms in the language of the reduced theory using the language of the reducing theory. Although discussion of general possibilities of reduction often drew on examples from physics, psychology, and social science as test cases, biology provides an especially good domain on which to focus the arguments. For, within molecular genetics, we have a well-established and articulated body of doctrine that uses the language of biochemistry to bear on issues that were previously tackled within classical genetics (the genetics descending from Mendel, worked out by Morgan and his associates in the early decades of the twentieth century). Instead of suggestions about how some vaguely characterized area of psychology might relate to some unknown future piece of neuroscience, we can look at the details of work in pre-molecular genetics and at the ways in which contemporary biologists have transformed it through the use of concepts and principles from biochemistry.

It was quickly apparent that Nagel's model couldn't apply without modification, but philosophers of biology drew different conclusions. In an important article, Kenneth Schaffner (1969) suggested that a modification of Nagel's approach would support the claim that classical genetics is reducible to molecular biology. David Hull (1972; 1974, ch. 1) argued that the reductionist claim was more deeply problematic, and that it represented another misguided effort to force scientific practice into philosophical preconceptions. In two articles (Kitcher 1982*a*, 1984*b*), I developed Hull's critique, proposing that the classical account of reduction failed for

three reasons: first, because the practices of classical and molecular genetics don't fit the conception of theory presupposed by the Nagel model; secondly, because main concepts of classical genetics, such as *gene*, cannot be specified in purely biochemical terms (that is, the pertinent 'bridge laws' aren't available; thirdly, because even if a derivation of some 'law' of genetics from principles of molecular biology were available, that derivation would fail to be explanatory (this point is amplified further in Kitcher 1999). Yet it seemed to me important to provide a framework in which the actual connections between molecular and classical genetics could be made clear, without relying on any concept of reduction. I proposed that the explanatory strategies of classical and molecular genetics are related in that molecular explanations deepen, or extend, those offered by the classical approach, and, in particular, that they allow for the analysis of relations between specific genotypes and specific phenotypes (in particular environments, of course). Kenneth Schaffner has also pursued a project of showing how the classical and molecular practices interrelate, although his preferred account remains closer to the logical empiricist conception of theories and theoretical reduction (Schaffner 1969, 1993).

Although many writers have supposed that attempts to reduce biology to physics and chemistry are thoroughly misguided (see e.g. Dupré 1993), anti-reductionism hasn't gone unquestioned. Alexander Rosenberg (1994) and Kenneth Waters (1990, 1994) have offered probing criticisms of the claims that classical concepts aren't specifiable in biochemical terms and that biochemical derivations wouldn't be explanatory. Similarly, Sahotra Sarkar (1998), while sceptical of reductionist claims, has offered a novel model of reductionism within which to situate the debate.

Unfortunately, attempts to go beyond the debate about reductionism to make philosophical sense of the view of life that emerges from contemporary molecular biology are relatively rare. Kenneth Schaffner has explored some of the ways in which explanations in biomedicine work (Schaffner 1993). William Bechtel and Robert Richardson consider a variety of lines of biological research in their study of complex sciences (Bechtel and Richardson 1992). There are also some discussions of experimentation and theory change in molecular biology (Culp 1995; Culp and Kitcher 1989). These endeavours tend to be driven by broader philosophical concerns—questions about scientific explanation, theory change, or the character of emergent properties. There is, however, at least one attempt to view molecular biology as generating an important new puzzle—a pioneering attempt to address the notion of biological information (Rosenberg 1985, ch. 8).

Ironically, this last issue has recently surfaced in philosophy of biology as a side consequence of evolutionary discussions. As I have already noted, some philosophers of biology have taken seriously the idea that there needs to be a new synthesis between evolutionary and developmental biology. A recent article by the evolutionary biologist John Maynard Smith on the notion of information in biology (Maynard Smith 2000) attracted commentary from several philosophers who are sceptical of the gene-centred perspective that they take to be standard in evolutionary

studies (Sarkar 2000; Godfrey-Smith 2000). Independently of how we formulate the theory of evolution, or of how we integrate evolution and development, however, contemporary molecular biology itself raises questions about how to understand the notion of information—which, according to the 'Central Dogma', is supposed to be able to flow from DNA to RNA, and from RNA to proteins, without being able to flow in the reverse directions.[6] Hence, I suggest, there was already a philosophically interesting question (spotted by Rosenberg in 1985), that didn't need the connection with evolution to make it worth pursuing.

Indeed, philosophical concern with an 'evolutionary-developmental synthesis' might better be directed at the prior clarification of developmental biology. During the past decades, molecular approaches to early embryology and to some aspects of development have made enormous strides—as in the work of Christiane Nüsslein-Volhard and her associates (which deservedly won the Nobel Prize in 1996). Providing a synthetic overview of what has already been accomplished would be a serious and important project in the philosophy of biology, one that would require careful delineation of suggestive (but imprecise) concepts of developmental stages and of levels of causation, and it seems evident to me that this project is a necessary precursor to fitting development into evolutionary biology. Further, the detailed molecular analyses raise intriguing questions about how to relate them to more general approaches to development, for example those that try to reconstruct developmental 'software' (Meinhardt 1998; Murray 1989; see Kitcher 1999).

I suspect that we are still only in the middle of the revolution begun in the 1940s with the birth of molecular biology, and the philosophical enterprises just reviewed only scratch the surface of the transformation that has so far occurred. The Human Genome Project has inspired some philosophers to look more closely at molecular biology, although the main discussions of that project centre, quite understandably, on its social implications. I'll now turn from my plea for a molecular turn in philosophical studies to a type of philosophical investigation that cuts across the categories that have been prominent so far, to take a look at philosophical appraisals of socially significant biology.

5. The Uses and Abuses of Biology

During the past decades, prominent biologists and social theorists have sometimes claimed that advances in the life sciences hold dramatic implications for our

[6] Since the discovery of retroviruses, it has been appreciated that information can sometimes flow from RNA to DNA, so that a more careful formulation of the Dogma is needed.

understanding of ourselves and our society. In the late 1960s, for example, investigations in behavioural genetics were advertised as showing the existence of a strong hereditary component in human intelligence. That conclusion seemed to have particular relevance in the context of test results revealing a fifteen-point average difference between the scores obtained on IQ tests by American blacks and Americans of Caucasian descent. It was hardly surprising that the title question of Arthur Jensen's famous article asked how much we can boost IQ (Jensen 1969).

The argument advanced by Jensen (and by Richard Herrnstein) began from the premiss that IQ tests provide a reliable measure of intelligence, independently of cultural background. It proceeds by offering estimates of the *heritability* of IQ test performance, taking these to be around 50 per cent. From this, Jensen and his followers draw the conclusion that there is a substantial genetic contribution to intelligence, and, as a result, efforts to modify the environment to raise the average in the black population to that in the white population are doomed to fail.

As Richard Lewontin pointed out in an early critique of this argument (1974*b*), there seems to be a misunderstanding of the notion of heritability. That notion is part of the technical apparatus of quantitative genetics, and is defined as follows: for any trait that admits of a quantitative measure, the heritability is the ratio of the variance due to genotype to the total variance. Plainly, then, heritability is a *population* statistic: in a population where there's no variance in environment, the heritability of any trait will be 1, whereas in a population in which there's no variance in genotype the heritability will be 0. These population-level features are quite independent of the question whether the trait in question is under stringent genetic control. Building on Lewontin's argument, the philosophers Ned Block and Gerald Dworkin presented a wide-ranging analysis of the flaws in the hereditarian argument (Block and Dworkin 1974). That analysis was later supplemented by Leon Kamin's demonstration that some of the data alleged to support high heritability estimates had been fudged (Kamin 1974), and by Gould's subsequent researches on the ways in which historical uses of IQ tests had been insensitive to cultural differences (Gould 1981).

That ended one round of the IQ debate, but, as so often, bad old arguments have amazing powers of regeneration. In the 1990s Richard Herrnstein, in collaboration with the social theorist Charles Murray, published a widely read book (Herrnstein and Murray 1994). Although the authors appeared to absorb the fundamental point made by Lewontin (and extended in the subsequent discussion), they claimed to reach similar conclusions to those endorsed earlier. A penetrating review by Block (1995) refashioned his original argument to the new context, and a subsequent article by Clark Glymour presented the Herrnstein–Murray claims within a general methodological framework (Glymour 1998).

I doubt that this will be the end of the matter. Contemporary behavioural genetics has crafted new molecular tools for attempting to understand the ways in which human behaviour is constrained by our genotypes. Although the Lewontin–Block

diagnosis shows clearly that heritability estimates alone will not provide information about the extent to which a trait can be modified by altering the environment, it is quite probable that the new tools (either alone or in combination with estimates of heritability) will be deployed to undergird the old conclusions. The idea of genetic determination seems endlessly fascinating. One main contribution of Lewontin and Block is to show us how to think clearly about that idea: for any given genotype, we can envisage a graph that shows the way in which a trait of social concern varies as the environment changes; the simplest genetic determinist theme is to suppose that the variation is relatively slight across the range of environments that we'd consider suitable for members of our species (for other themes, see Kitcher 2000). The challenge for people who think that our genes set limits to productive social policies is to amass evidence that enables them to draw the appropriate graph, and to show that it accords with their dismal predictions about the possibilities. Critics have to scrutinize carefully the methods that are used, and to identify the points (if any) at which unwarranted conclusions are drawn. That requires constant attention to the ways in which study of the genetics of behaviour evolves. Although it's easy to sympathize with those (like Susan Oyama) who yearn to solve nature–nurture problems at one stroke (by the 'stake-in-the-heart move'), there's no substitute for piecemeal consideration of the latest arguments.[7]

The same holds for another recurrent controversy of past decades, one that began with popular work in human ethology, became much more visible in the human sociobiology of the 1970s and 1980s, and that has metamorphosed into claims and counter-claims of contemporary evolutionary psychology. In all these instances, enthusiasts suggest that understanding the evolutionary past of our species will help us refine our views about who we are, how our minds work, and what we can aspire to be. I'll illustrate the programme by starting with one of the most celebrated instances, E. O. Wilson's Sociobiology (Wilson 1975, 1978).

A distinguished entomologist, famous for his ground-breaking work on social insects, Wilson was impressed by the theoretical developments of the 1960s, and offered a wide-ranging survey of the evolution of sociality across a wide range of species. Applying his favoured tools to human beings, he defended a biological account of behavioural differences between the sexes (both in propensities to various kinds of work and in sexual responses), of xenophobia, of tendencies to aggression, and of the hierarchical structure of human society. Part of the argument about sexual behaviour can serve as an example. According to Wilson, asymmetries in numbers and size of gametes (men produce a lot of sperm, women a far smaller number of eggs), coupled with greater female 'investment' in progeny at the embryonic stages, serve as the basis for selection pressures that can be expected to favour

[7] So far, philosophers have not paid much attention to the sophisticated work now being done in molecular behavioural genetics, although the recent work of Kenneth Schaffner (1993, 1998) is a notable contribution.

different sexual attitudes, men being relatively hasty and promiscuous, women relatively coy. We can thus expect that the differences we observe in male–female behaviour are traceable to underlying genes, and that there's little that can be done to modify these differences (1975, 1978).

Although some philosophers were more impressed by this style of argument (Ruse 1979), it quickly attracted detailed objections. Critics pointed out that evolutionary expectations depend on the provision of detailed models that show how fitness is affected by various factors, that they are also subject to conditions about the ability of genetic variations to modify particular manifest traits (a point that leads into the controversy about adaptationism), that human behaviour has to be understood as responsive to cultural transmission, and, perhaps most importantly, that even granting a genetic basis for a trait, nothing follows about the extent to which the trait is now changeable by altering the environment (Kitcher 1985*b*; Lewontin *et al*. 1984).[8] These objections are quite consistent with a positive view of some ventures in sociobiology, for example careful studies of animals that combine field observations with attention to genetic and ecological models and that draw no deterministic conclusions; indeed, Wilson's own work on the social insects can be seen as far more rigorous and cautious than his speculations about our species.

During the late 1980s human sociobiology became almost invisible. A number of researchers continued to insist that a Darwinian perspective could inform anthropology, but this was typically qualified with the recognition that the pitfalls of human sociobiology should be avoided. In the past decade or so, however, more of the old themes have re-emerged. Authors have proposed that evolutionary theory can offer insights into the character of human psychology (Barkow *et al*. 1992), ranging from our sensitivity to social cheating (Cosmides and Tooby 1992) to the springs of sexual desire (Buss 1994, 2000) and rape (Thornhill and Palmer 2000). The worst excesses of this literature recapitulate the errors diagnosed earlier, and philosophers, as well as scientists, have begun to develop critiques (Cheng and Holyoak 1989; Lloyd 1999; Dupré 2001; Travis 2002).

As with the IQ controversy, it would be desirable to settle matters once for all, but there's no substitute for piecemeal analyses. For the problem doesn't stem from some systematic flaw in evolutionary theory but from the ways in which perfectly good evolutionary tools are applied. Wilson drew on the same arsenal of techniques in his work on ants and in his claims about human behaviour; the difference lay in the care and caution with which those techniques were applied.

Besides focusing on the forms of behaviour just considered, evolutionary studies of human beings are also pertinent to areas of traditional philosophical concern. Since the late nineteenth century, many thinkers have pondered the connection

[8] Here we re-encounter a point made in discussing the IQ controversy. Simply supposing that, in some environment, there's a causal connection between a genotype and a trait doesn't tell us whether that connection would also obtain in other environments.

(if any) between evolution and ethics. Human sociobiology took a forthright stand on this issue, claiming that the content of 'ethical imperatives' could be derived from an understanding of our evolutionary heritage. When the claim was made more concrete, it typically amounted either to the idea that we have a moral obligation to promote the replication of human DNA (Wilson 1978) or to moral prohibitions against such things as incest (Ruse and Wilson 1986). It's not hard to show that such simple connections are suspect (Kitcher 1992). Fortunately, some scholars have found more subtle ways of linking evolutionary ideas to our moral and political concerns. Elliott Sober and David Sloan Wilson use their account of the evolution of altruism (in the biologist's sense) as a prelude to giving an account of human psychological altruism and to exploring the possibility that psychological altruism might have evolved under natural selection (Sober and Wilson 1998). Brian Skyrms has drawn on the techniques of evolutionary game theory (Maynard Smith 1982) to show how we can make sense of elementary features of social and political arrangements (Skyrms 1996). These ventures evade my earlier objections because they make precise use of evolutionary concepts and methods, and they are appropriately restrained in drawing their conclusions—in many instances, Sober and Wilson, and Skyrms, are interested in showing how a particular outcome is possible rather than trying to infer something 'deep' about genetic determination of behaviour. From a different direction, philosophers interested in moral theory have found inspiration in evolutionary discussions of altruism and cooperation, and have tried to make biological links with meta-ethical positions (non-cognitivism in Gibbard 1990; Humean expressivism in Blackburn 2000). Although the history of speculations about evolution and ethics isn't encouraging, we may finally have reached a stage in which careful research in this area will bear dividends.

Another important locus of philosophical discussion is the relationship between evolutionary biology and religion (particularly Christianity). Despite the fact that the Church of England made its peace with Darwin in 1882—he was, after all, buried in Westminster Abbey, against the wishes of his family—there have been many people, especially in North America, who have viewed evolutionary biology as antipathetic to religion. In the 1970s and early 1980s such people obtained enough popular support to inspire many authors, including philosophers, to defend evolution against the critiques of fundamentalist 'creation science' (Ruse 1982; Futuyma 1983; Kitcher 1982*b*). More recently, creationists have articulated a more sophisticated anti-evolutionary position, one no longer committed to the literal truth of Genesis, emphasizing the role of 'intelligent design' (Behe 1996; Johnson 1993; Dembski 1998). In response, a number of philosophers have dissected the arguments supposed to show the presence of design in the universe and the corresponding limitation of orthodox Darwinism (Pennock 1999, 2002).

But the major issue of the compatibility of Darwinism and Christianity remains. A popular approach, defended in (Kitcher 1982*b*; Gould 2000; Ruse 2001), is to insist that enlightened religious believers can slough off those parts of religious

texts for which evolutionary biology causes trouble and preserve the central moral messages. Despite my own earlier advocacy of this compatibilism, it no longer appears so easy. For, as Darwin saw, and as Dawkins has forcefully emphasized, a standard Darwinian view of the history of life exacerbates the traditional problem of evil—Dawkins puts the point by asking us to reconstruct the deity's utility function from the observed phenomena (Dawkins 1995). In my own view, the difficulties of combining Darwinism with religious belief result from a much more general war between various sciences and religion, one that also involves historical reconstructions of major religious texts and of the growth of religious belief, anthropological studies of religious diversity, psychological investigations of the causes of 'religious experiences', and philosophical dissection of the concept of faith. In effect the war is fought on many fronts and although religion's losses in Darwinian battles are extremely severe, the real trouble is that the believer is hard-pressed everywhere (Kitcher, forthcoming).

I'll conclude with a quick look at another socially relevant area of biology, the current research on genome sequencing, with our own species as a special case. The advertisements for the Human Genome Project promise that investment in this research will offer cures for major diseases and disabilities, and, indeed, we would be extremely unlucky if we were not *eventually* to be able to do better with some of the chronic conditions to which major efforts are currently directed (cancer, heart disease, diabetes). In the short term, however, the immediate applications of the power to obtain DNA sequences are likely to lie in techniques of identification (used already to liberate innocent people from prisons) and most of all in predictive testing. The uses of predictive tests have now been thoroughly debated (Holtzmann 1989; Nelkin and Tancredi 1994; Hubbard and Wald 1993; Andrews *et al.* 1994). It's far from clear, however, that the affluent societies in which tests are likely to be available in greatest profusion are yet equipped with the social mechanisms to ensure that people are adequately protected. Already in the United States those testing positive for genetic conditions have found their lives disrupted by loss of jobs and insurance, and this is likely to increase in coming decades as the power to test grows. Moreover, matters will become more complicated with the application of molecular genetics to questions about human behaviour (as noted in Section 5 above), and several authors have recognized the possibility of a new form of eugenics, one that might in principle be benign but that is likely in practice to recapitulate old errors (Duster 1990; Kitcher 1996).

Philosophical work on these social issues ranges from relatively abstract considerations in moral theory (Heyd 1992) to more detailed involvement with the possibilities furnished by contemporary biology (Harris 1992; Kitcher 1996; Buchanan *et al.* 2000). Some issues, especially the threats and promises of molecular behavioural genetics, need to be explored more thoroughly than has been done so far. There is also a broader question about the concentration of biomedical research on the diseases that afflict citizens of affluent nations, especially when the sequencing

techniques provide opportunities for developing vaccines that might alleviate the misery of millions in the developing world. A healthy outgrowth of philosophical concern with the Human Genome Project—and socially significant biology more generally—might be a more resolute attempt to pose and answer ethical and political questions about scientific research.

6. An Apologetic Conclusion

Although I have swept through many areas of recent discussion at a brisk pace, there are other questions that are being (or ought to be) addressed, questions that I have neglected. I'll close with a very brief mention of some of these.

The notion of a law of nature was central to logical empiricist philosophy of science, and, since the 1970s, a number of philosophers have attempted to characterize laws without imposing on themselves the Humean scruples that logical empiricists tried to honour. Philosophy of biology has contributed to this debate both by raising questions about the contingency of laws (Beatty 1995), and by reviving the controversy about the sense of "law" in which biology can lay claim to laws of its own (Mitchell 2000). Here, as in other examples I've discussed above, we can recognize the impact of the philosophy of biology on general philosophy of science.

There have also been consequences for other areas of philosophy. Discussions in the philosophy of mind have been transformed by a richer understanding of neurobiology, stemming from Patricia Churchland's pioneering (1985) and more recent work by Kathleen Akins (1996) and Brian Keeley (2002). Within the philosophy of language, attempts to provide a naturalistic account of semantics (and of mental representation) have drawn on conceptions of biological function and on research on animal communication (Dretske 1988; Millikan 1984; Godfrey-Smith 1996).

But some areas of biology that seem to call for philosophical attention have been strangely neglected. Very little has been done to clarify the notion of biodiversity and to elaborate a philosophical foundation for conservation. Although a few philosophers have contributed to debates about artificial life (Boden 1996), this is another area in which there is abundant philosophical work to be done. Yet, as I've insisted in Section 5, the major gap in contemporary philosophy of biology is the failure to come to terms with the many facets of molecular biology, and, in particular, its transformation of physiology, cell biology, and developmental biology.

It would be wrong to end on a note of complaint. The last thirty years have witnessed so many diverse and fruitful interactions between philosophy and biology that it has become impossible for any philosopher of science (perhaps any philosopher) to write in ignorance of the main concepts and themes of the life sciences.

Philosophy owes a debt to the pioneers who saw the importance of biological research. I have tried to provide a sketch of the exciting enterprise they started.[9]

References

Akins, Kathleen (1996). 'Of Sensory Systems and the "Aboutness" of Mental States'. *Journal of Philosophy*, 93: 337–72.

Andrews, Lori, Fullerton, Jane E., Holtzmann, Neil A., and Motulsk, Arno G. (1994). *Assessing Genetic Risks*. Washington: National Academy Press.

Barkow, Jerome, Cosmides, Leda, and Tooby, John (eds.) (1992). *The Adapted Mind*. New York: Oxford University Press.

Beatty, John (1980*a*). 'Optimal-Design Models and the Strategy of Model-Building in Evolutionary Biology'. *Philosophy of Science*, 47: 532–61.

—— (1980*b*). 'What's Wrong with the Received View of Evolutionary Theory?', in Peter Asquith and Ronald Giere (eds.), *PSA 1980*, ii. East Lansing, Mich.: Philosophy of Science Association.

—— (1995). 'The Evolutionary Contingency Thesis', in G. Walters and J. Lennox (eds.), *Concepts, Theories, and Rationality in the Biological Sciences*. Pittsburgh: University of Pittsburgh Press.

Bechtel, William, and Richardson, Robert (1992). *Discovering Complexity*. Princeton: Princeton University Press.

Beckner, Morton (1959). *The Biological Way of Thought*. New York: Columbia University Press.

Behe, Michael (1996). *Darwin's Black Box*. New York: Free Press.

Beurton, Peter, Falk, Raphael, and Reinberger, Hans-Jörg (eds.) (2000). *The Concept of the Gene in Development and Evolution*. Cambridge: Cambridge University Press.

Bigelow, John, and Pargetter, Robert (1987). 'Functions'. *Journal of Philosophy*, 84: 181–96.

Blackburn, Simon (2000) *Ruling Passions*. New York: Oxford University Press.

Block, N. J., and Dworkin, Gerald (1974). 'IQ, Heritability and Inequality'. *Philosophy and Public Affairs*, 4: 1–99. Repr. in N. J. Block and Gerald Dworkin (eds.), *The I.Q. Controversy*. New York: Pantheon, 1976.

—— (1995). 'How Heritability Misleads About Race'. *Cognition*, 56: 99–126.

Boden, Margaret (ed.) (1996). *The Philosophy of Artificial Life*. Oxford: Oxford University Press.

Boorse, Christopher (1976). 'Wright on Functions'. *Philosophical Review*, 85: 70–86.

Brandon, Robert (1984). 'The Levels of Selection', in Robert Brandon and Richard Burian (eds.), *Genes, Organisms, Populations*. Cambridge, Mass.: MIT Press.

—— (1990). *Adaptation and Environment*. Princeton: Princeton University Press.

Buchanan, Allen, Brock, Dan, Daniels, Norman, and Wikler, Daniel (2000). *From Chance to Choice*. Cambridge: Cambridge University Press.

Buss, David (1994). *The Evolution of Desire*. New York: Basic Books.

—— (2000). *The Dangerous Passion*. New York: Free Press.

[9] Many thanks to Frank Jackson and Michael Smith, for their comments on an earlier version, and, above all, for their patience.

Cheng, Patricia, and Holyoak, Keith (1989). 'On the Natural Selection of Reasoning Theories'. *Cognition*, 33: 285–313.
Churchland, Patricia (1985). *Neurophilosophy*. Cambridge, Mass.: MIT Press.
Cosmides, Leda, and Tooby, John (1992). 'Cognitive Adaptations for Social Exchange', in J. Barkow, L. Cosmides, and J. Tooby (eds.), *The Adapted Mind*. New York: Oxford University Press.
Culp, Sylvia (1995). 'Objectivity in Experimental Inquiry: Breaking Data-Technique Circles', *Philosophy of Science*, 62: 430–50.
—— and Kitcher, Philip (1989). 'Theory Structure and Theory Change in Contemporary Molecular Biology'. *British Journal for the Philosophy of Science*, 40: 459–83.
Cummins, Robert (1973). 'Functional Analysis'. *Journal of Philosophy*, 72: 741–64.
Darwin, Charles (1859). *The Origin of Species*. London: John Murray.
Dawkins, Richard (1976). *The Selfish Gene*. Oxford: Oxford University Press.
—— (1982). *The Extended Phenotype*. San Francisco: Freeman.
—— (1987). *The Blind Watchmaker*. London: Longmans.
—— (1995). *River Out of Eden*. New York: Basic Books.
Dembski, William (1998). *The Design Inference*. Cambridge: Cambridge University Press.
Dennett, Daniel (1995). *Darwin's Dangerous Idea*. New York: Simon & Schuster.
Dretske, Fred (1988). *Explaining Behavior*. Cambridge, Mass.: MIT Press.
Dupré, John (1981). 'Natural Kinds and Biological Taxa'. *Philosophical Review*, 90: 66–90.
—— (ed.) (1987). *The Latest on the Best*. Cambridge, Mass.: MIT Press.
—— (1993). *The Disorder of Things*. Cambridge, Mass.: Harvard University Press.
—— (2001). *Human Nature and the Limits of Science*. Oxford: Oxford University Press.
Duster, Troy (1990). *Backdoor to Eugenics*. New York: Routledge.
Eldredge, Niles (1985). *Unfinished Synthesis*. New York: Oxford University Press.
Ereshevsky, Marc (ed.) (1992). *The Units of Evolution*. Cambridge, Mass.: MIT Press.
Futuyma, Douglas (1983). *Science on Trial*. New York: Pantheon.
Ghiselin, Michael (1974). 'A Radical Solution to the Species Problem'. *Systematic Zoology*, 23: 536–44.
Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, Mass.: Harvard University Press.
Glymour, Clark (1998). 'What Went Wrong?' *Philosophy of Science*, 65: 1–32.
Godfrey-Smith, Peter (1993). 'Functions: Consensus without Unity'. *Pacific Philosophical Quarterly*, 74: 196–208.
—— (1994). 'A Modern History Theory of Functions'. *Noûs*, 28: 344–62.
—— (1996). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.
—— (2000). 'On the Theoretical Role of "Genetic Coding" ', *Philosophy of Science*, 67: 26–44.
—— and Lewontin, Richard (1993). 'The Dimensions of Selection'. *Philosophy of Science*, 60: 373–95.
Goudge, T. A. (1961). *The Ascent of Life*. Toronto: University of Toronto Press.
Gould, Stephen Jay (1978). *Ontogeny and Phylogeny*. Cambridge, Mass.: Harvard University Press.
—— (1980*a*). 'Caring Groups and Selfish Genes', in Gould, *The Panda's Thumb*. New York: Norton.
—— (1980*b*). 'Is a New and General Theory of Evolution Emerging?' *Paleobiology*, 6: 119–30.
—— (1981). *The Mismeasure of Man*. New York: Norton.
—— (1982). 'Darwinism and the Expansion of Evolutionary Theory'. *Science*, 216: 380–7.

Gould, Stephen Jay (2000). *Rocks of Ages*. New York: Ballantine.
——(2002). *The Structure of Evolutionary Theory*. Cambridge, Mass.: Harvard University Press.
——and Lewontin, Richard C. (1979). 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme'. *Proceedings of the Royal Society of London*, ser. B, 205: 581–98.
——and Vrba, Elizabeth (1982). 'Exaptation: A Missing Term in the Science of Form'. *Paleobiology*, 8: 4–15.
Grene, Marjorie (1959). 'Two Evolutionary Theories'. *British Journal for the Philosophy of Science*, 9: 110–27, 185–93.
Griffiths, Paul, and Gray, Russell (1994). 'Developmental Systems and Evolutionary Explanation'. *Journal of Philosophy*, 91: 277–304.
Hamilton, W. D. (1964). 'The Genetical Evolution of Social Behavior', I and II. *Journal of Theoretical Biology*, 7: 1–16, 17–32.
Harris, John (1992). *Wonderwoman and Superman*. Oxford: Oxford University Press.
Hempel, C. G. (1965). 'The Logic of Functional Explanation', in C. G. Hempel, *Aspects of Scientific Explanation*. New York: Free Press.
Herrnstein, R., and Murray, C. (1994). *The Bell Curve*. New York: Free Press.
Heyd, David (1992). *Genethics*. Berkeley: University of California Press.
Ho, Mae-Wan, and Fox, Sidney (eds.) (1988). *Evolutionary Processes and Metaphors*. Chichester: John Wiley.
——and Saunders, Peter T. (1984). *Beyond Neo-Darwinism*. London: Academic Press.
Holtzmann, Neil A. (1989). *Proceed with Caution*. Baltimore: Johns Hopkins University Press).
Hubbard, Ruth, and Wald, Elijah (1993). *Exploding the Gene Myth*. Boston: Beacon.
Hull, David (1972). 'Reduction in Genetics—Biology or Philosophy?' *Philosophy of Science*, 39: 491–9.
——(1974). *Philosophy of Biological Science*. Englewood Cliffs, NJ: Prentice-Hall.
——(1976). 'Are Species Really Individuals?' *Systematic Zoology*, 25: 174–91.
——(1978). 'A Matter of Individuality'. *Philosophy of Science*, 45: 335–60.
——(1979). 'The Limits of Cladism'. *Systematic Zoology*, 28: 414–38.
——(1981). 'Units of Evolution: A Metaphysical Essay', in R. Jensen and R. Harré (eds.), *The Philosophy of Evolution*. Brighton: Harvester.
——(1988). *Science as a Process*. Chicago: University of Chicago Press.
——and Ruse, Michael (eds.) (1998). *Philosophy of Biology*. Oxford: Oxford University Press.
Jensen, A. R. (1969). 'How Much Can We Boost I.Q. and Scholastic Achievement?' *Harvard Educational Review*, 39: 1–123.
Johnson, Phillip (1993). *Darwin on Trial*. Washington: Regnery Gateway.
Kamin, Leon (1974). *The Science and Politics of IQ*. Potomac, Md.: Erlbaum.
Kauffman, Stuart (1993). *The Origins of Order*. New York: Oxford University Press.
Keeley, Brian (2002). 'Making Sense of the Senses'. *Journal of Philosophy*, 99: 5–28.
Kitcher, Philip (1982*a*). 'Genes'. *British Journal for the Philosophy of Science*, 33: 337–59.
——(1982*b*). *Abusing Science: The Case Against Creationism*. Cambridge, Mass.: MIT Press.
——(1984*a*). 'Species'. *Philosophy of Science*, 51: 308–33.
——(1984*b*). '1953 and All That: A Tale of Two Sciences'. *Philosophical Review*, 93: 335–73.
——(1985*a*). 'Darwin's Achievement', in N. Rescher (ed.), *Reason and Rationality in Science*. Washington: University Press of America.

——(1985*b*). *Vaulting Ambition*. Cambridge, Mass.: MIT Press.
——(1989). 'Some Puzzles About Species', in M. Ruse (ed.), *What the Philosophy of Biology Is*. Dordrecht: Kluwer.
——(1992). 'Four Ways of "Biologicizing" Ethics', in Kurt Bayertz (ed.), *Evolution und Ethik. Biologische Grundlagen der Moral?* Stuttgart: Reclam. Repr. in English in Elliott Sober (ed.), *Conceptual Issues in Evolutionary Theory*, 2nd edn. Cambridge, Mass.: MIT Press.
——(1993). 'Function and Design', in P. French, T. Uehling, and H. Wettstein (eds.), *Midwest Studies in Philosophy*, xviii: *Philosophy of Science*. Minneapolis: University of Minnesota Press.
——(1996). *The Lives to Come*. New York: Simon & Schuster.
——(1999). 'The Hegemony of Molecular Biology'. *Biology and Philosophy*, 14: 195–210.
——(2000). 'Battling the Undead: How (and How Not) to Resist Genetic Determinism', in Rama Singh, Costas Krimbas, Diane Paul, and John Beatty (eds.), *Thinking About Evolution: Historical, Philosophical and Political Perspectives*. Cambridge: Cambridge University Press.
——(forthcoming). 'The Many-Sided Conflict Between Science and Religion', in William Mann (ed.), *Companion to the Philosophy of Religion*. Oxford: Blackwell.
Levins, Richard, and Lewontin, Richard C. (1985). *The Dialectical Biologist*. Cambridge, Mass.: Harvard University Press.
Lewontin, Richard C. (1974*a*). *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.
——(1974*b*). 'The Analysis of Variance and the Analysis of Causes'. *American Journal of Human Genetics*, 26: 400–11.
——Rose, Steven, and Kamin, Leon (1984). *Not in our Genes*. New York: Pantheon.
Lloyd, Elisabeth (1983). 'The Nature of Darwin's Support for the Theory of Natural Selection'. *Philosophy of Science*, 50: 112–29.
——(1988). *The Structure and Confirmation of Evolutionary Theory*. Westport, Conn.: Greenwood Press.
——(1999). 'Evolutionary Psychology: The Burden of Proof'. *Biology and Philosophy*, 14: 211–33.
Maynard Smith, John (1964). 'Group Selection and Kin Selection'. *Nature*, 201: 1145–7.
——(1978). 'Optimization Theory in Evolution'. *Annual Review of Ecology and Systematics*, 9: 31–56.
——(1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
——(2000). 'The Concept of Information in Biology'. *Philosophy of Science*, 67: 177–94.
Mayr, Ernst (1942). *Systematics and the Origin of Species*. New York: Columbia University Press.
——(1963). *Animal Species and Evolution*. Cambridge, Mass.: Harvard University Press.
——(1983). 'How to Carry Out the Adaptationist Program?' *American Naturalist*, 121: 324–33.
Meinhardt, Hans (1998). *The Algorithmic Beauty of Sea Shells*. New York: Springer.
Millikan, Ruth (1984). *Language, Thought, and Other Biological Categories*. Cambridge, Mass.: MIT Press.
Mills, Susan, and Beatty, John (1979). 'The Propensity Interpretation of Fitness'. *Philosophy of Science*, 46: 263–86.
Mitchell, Sandra (1993). 'Dispositions or Etiologies?' *Journal of Philosophy*, 90: 249–59.
——(1995). 'Function, Fitness and Disposition'. *Biology and Philosophy*, 10: 39–54.

Mitchell, Sandra (2000). 'Dimensions of Scientific Law'. *Philosophy of Science*, 67: 242–65.
Murray, John D. (1989). *Mathematical Biology*. New York: Springer.
Nagel, Ernest (1962). *The Structure of Science*. London: Routledge.
—— (1979). *Teleology Revisited*. New York: Columbia University Press.
Neander, Karen (1991). 'Functions as Selected Effects'. *Philosophy of Science*, 58: 168–84.
Nelkin, Dorothy, and Tancredi, Laurence (1994). *Dangerous Diagnostics*. Chicago: University of Chicago Press.
Orzack, S., and Sober, Elliott (1994). 'Optimality Methods and the Test of Adaptationism'. *American Naturalist*, 143: 361–80.
Oster, George, and Wilson, E. O. (1978). *Caste and Ecology in the Social Insects*. Princeton: Princeton University Press.
Oyama, Susan (1985). *The Ontogeny of Information*. Cambridge: Cambridge University Press.
Pennock, Robert (1999). *Tower of Babel*. Cambridge, Mass.: MIT Press.
—— (ed.) (2002). *Intelligent Design Creationism and its Critics*. Cambridge, Mass.: MIT Press.
Raff, R. (1996). *The Shape of Life*. Chicago: University of Chicago Press.
Rosenberg, Alexander (1982). 'On the Propensity Definition of Fitness'. *Philosophy of Science*, 49: 268–73.
—— (1983). 'Fitness'. *Journal of Philosophy*, 80: 457–73.
—— (1985). *The Structure of Biological Science*. Cambridge: Cambridge University Press.
—— (1994). *Instrumental Biology and the Disunity of Science*. Chicago: University of Chicago Press.
Ruse, Michael (1979). *Sociobiology: Sense or Nonsense?* Dordrecht: Reidel.
—— (1982). *Darwinism Defended*. Reading, Mass.: Addison-Wesley.
—— (2001). *Can a Darwinian Be a Christian?* Cambridge: Cambridge University Press.
—— and Wilson, E. O. (1986). 'Moral Philosophy as Applied Science'. *Philosophy*, 61: 173–92.
Sarkar, Sahotra (1998). *Genetics and Reductionism*. Cambridge: Cambridge University Press.
—— (2000). 'Information in Genetics and Developmental Biology'. *Philosophy of Science*, 67: 208–13.
Schaffner, Kenneth (1969). 'The Watson–Crick Model and Reductionism'. *British Journal for the Philosophy of Science*, 20: 325–48.
—— (1993). *Discovery and Explanation in Biology and Medicine*. Chicago: University of Chicago Press.
—— (1998). 'Genes, Behavior, and Developmental Emergentism: One Process, Indivisible', *Philosophy of Science*, 65: 209–52.
Skyrms, Brian (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Sober, Elliott (1984). *The Nature of Selection*. Cambridge Mass.: MIT Press.
—— (1988). *Reconstructing the Past*. Cambridge, Mass.: MIT Press.
—— (1998). 'Six Sayings About Adaptationism', in David Hull and Michael Ruse (eds.), *Philosophy of Biology*. Oxford: Oxford University Press.
—— and Lewontin, Richard (1982). 'Artifact, Cause and Genic Selection'. *Philosophy of Science*, 49: 157–80.
—— and Wilson, David Sloan (1998). *Unto Others*. Cambridge, Mass.: Harvard University Press.
Splitter, Laurence (1988). 'Species and Identity'. *Philosophy of Science*, 55: 323–48.

Sterelny, Kim, and Kitcher, Philip (1988). 'The Return of the Gene'. *Journal of Philosophy*, 85: 339–60.
—— and Griffiths, Paul (1999). *Sex and Death*. Chicago: University of Chicago Press.
Thompson, R. P. (1983). 'The Structure of Evolutionary Theory: A Semantic Approach'. *Studies in the History and Philosophy of Science*, 14: 215–29.
Thornhill, Randy, and Palmer, Craig (2000). *The Natural History of Rape*. Cambridge, Mass.: MIT Press.
Tinbergen, Niko (1963). 'On Aims and Methods of Ethology'. *Zeitschrift für Tierpsychologie*, 20: 410–33.
Travis, Cheryl (ed.) (2002). *Evolution, Gender, and Violence*. Cambridge, Mass.: MIT Press.
Waters, C. Kenneth (1990). 'Why the Anti-Reductionist Consensus Won't Survive: The Case of Classical Genetics', in Arthur Fine, Micky Forbes, and Linda Wessels (eds.), *PSA 1990*, i. East Lansing, Mich.: Philosophy of Science Association.
—— (1991). 'Tempered Realism About the Force of Selection'. *Philosophy of Science*, 58: 553–73.
—— (1994). 'Genes Made Molecular'. *Philosophy of Science*, 61: 163–85.
Williams, George C. (1966). *Adaptation and Natural Selection*. Princeton: Princeton University Press.
Williams, Mary (1970). 'Deducing the Consequences of Evolution: A Mathematical Model'. *Journal of Theoretical Biology*, 29: 343–85.
Wilson, E. O. (1975). *Sociobiology*. Cambridge, Mass.: Harvard University Press.
—— (1978). *On Human Nature*. Cambridge, Mass.: Harvard University Press.
Wimsatt, William (1981). 'The Units of Selection and the Structure of the Multi-Level Genome', in Peter Asquith and Ronald Giere (eds.), *PSA 1980*, ii. East Lansing, Mich.: Philosophy of Science Association.
—— (1986). 'Developmental Constraints, Generative Entrenchment, and the Innate–Acquired Distinction', in W. Bechtel (ed.), *Integrating Scientific Disciplines*. Amsterdam: Nijhoff.
Woodger, J. H. (1937). *The Axiomatic Method in Biology*. Cambridge: Cambridge University Press.
Wright, Larry (1973). 'Functions'. *Philosophical Review*, 82: 139–68.