Butler on Virtue, Self-Interest and Human Nature

Ralph Wedgwood

In his Sermons, Joseph Butler argued for a series of extraordinarily subtle and perceptive claims about the relations between virtue and self-interest. Unfortunately, there has been a great deal of controversy among Butler’s interpreters about what exactly these claims amount to, and about what role these claims play in the overall project of his Sermons. In this essay, I shall set out and defend a new interpretation of Butler’s argument. Although I shall argue that in the end, Butler’s argument is not completely successful, I hope that my interpretation will make it plausible that Butler’s argument is both more distinctive and original, and also more defensible, than most commentators have supposed.

1. Butler’s “naturalist” project

Butler announces his project at the very beginning of the Preface. It is to answer “the important question, What is the rule of life?” (P 1). In particular, his answer is that we have “obligations to the practice of virtue” (P 12). By this he seems to mean that we have overriding reasons to live virtuously and to comply with the requirements of morality. However, the ultimate aim behind Butler’s project is not philosophical at all. He is a

---

* An earlier draft of this paper was presented to an audience at the University of Reading. I am grateful to members of that audience, and also to Paul Bloomfield, the editor of this volume, to Stephen Darwall, and to my Oxford colleagues Robert Adams, Bill Child, Antony Eagle, David Charles, Dorothy Edgington, and Oliver Pooley, for helpful comments on earlier drafts.

1 References to the Fifteen Sermons are to Sermon and paragraph number, according to The Works of Bishop Butler, ed. J. H. Bernard (London: The English Theological Library, 1900). ‘P’ refers to the Preface to the Fifteen Sermons.
preacher, and his arguments are sermons. Thus, his ultimate aim is homiletic and therefore pastoral. He argues that we have “obligations to the practice of virtue” as a way of exercising spiritual care for his congregation, by strengthening their disposition to lead a virtuous life.

Butler believes that he will be following a distinctive method to argue for this answer to the question. As he says (P 12):

There are two ways in which the subject of morals may be treated. One begins from inquiring into the abstract relations of things; the other from a matter of fact, namely, what the particular nature of man is, its several parts, their economy or constitution; from whence it proceeds to determine what course of life it is, which is correspondent to this whole nature.

These two methods, he says, “both lead to the same thing, our obligations to the practice of virtue; and thus they exceedingly reinforce and strengthen each other” (P 12). But as he goes on to explain, in his Sermons he is chiefly following the second of these two methods: “The following discourses proceed chiefly in this latter method. The three first wholly” (P 13).

Commentators generally agree that the first method is the rationalist method, which Butler almost certainly associated with the work of Samuel Clarke and William Wollaston. The characteristic feature of this rationalist method is that it seeks to discover necessary truths by means of a priori reflection. By contrast, the second method starts out from contingent facts that are known on the basis of empirical observation. So Butler is claiming that the arguments of his Sermons are largely based on the contingent facts of empirical observation. This is not, however, because he is an empiricist who rejects the rationalist

---

2 In his youth Butler had corresponded with Clarke; see the letters in Vol. 1 of The Works of Joseph Butler, pp. 311–339. Butler refers to William Wollaston as “a late author of great and deserved reputation” in the Preface to the Sermons (P 13).

3 Compare I.6, note: “whether man be thus, or otherwise constituted, what is the inward frame in this particular, is a mere question of fact or natural history, not proveable immediately by reason. It is therefore to be judged of and determined in the same way other
method; on the contrary, he explicitly accepts the validity of the rationalist method. He follows the empirical method simply because it better suits his homiletic purposes; as he says, the empirical method “is in a peculiar manner adapted to satisfy a fair mind, and is more easily applicable to the several particular relations and circumstances of life” (P 12).

Specifically, Butler’s empirical method is based on an inquiry into human nature. Here it becomes crucial to understand exactly what Butler meant by speaking of human nature. One good way to understand what Butler means is to look at the earlier works of moral philosophy that are clearly influencing him.

First, Butler explicitly claims that in saying that virtue consists in following nature, he is repeating the view of “the ancient moralists” (P 13). He seems to be thinking chiefly of the ancient Stoics here. When he says that the ancients had some “inward feeling … which they chose to express in this manner, that man is born to virtue, that it consists in following nature, and that vice is more contrary to this nature than tortures or death”, he is giving a close paraphrase of a passage where Cicero presents the Stoic view (De Officiis iii.21). So in appealing to human nature in this way, Butler takes himself to be following these ancient Stoic philosophers.4

At the same time, Butler was obviously well acquainted with Shaftesbury’s Inquiry

---

4 Besides Cicero, the only other obvious allusions to ancient authors in Butler’s ethical works are in the “Dissertation Upon the Nature of Virtue” (The Analogy of Religion, Appendix II), in The Works of Bishop Butler, vol. 2. Here too it is Stoic philosophers that he refers to — namely, Epictetus (§ 1, note) and Marcus Aurelius (§ 2, note). On Butler’s use of Stoic ideas, see especially Terence Irwin, “Stoic Naturalism in Butler”, in Hellenistic and Early Modern Philosophy, ed. Jon Miller (Cambridge University Press, 2003).
concerning Virtue or Merit, which he refers to explicitly in the Preface (P 26). Shaftesbury also seems to have been profoundly influenced by some of the ancients (including the Stoics Epictetus and Marcus Aurelius). Moreover, Shaftesbury bases his inquiry into virtue on an account of the “Constitution and Frame of Nature”, and in general he believes that nature is known “by Study and Observation”. Moreover, Shaftesbury’s work seems clearly to appeal to empirical observations of human nature, and largely to lack the attempts at formal demonstrative reasoning that are such a prominent feature of such works as Clarke’s Discourse concerning the Unchangeable Obligations of Natural Religion. For these reasons, then, it seems plausible that Butler would also have regarded Shaftesbury as also employing this second method. This is not to say that Butler is just a slavish follower of Shaftesbury. Far from it. Butler clearly believes that his execution of this naturalist empirical method avoids a certain crucial “material deficiency or omission” that mars Shaftesbury’s approach (P 26). But nonetheless at bottom it is the same method that both philosophers are employing.

Now it should be quite uncontroversial that both Shaftesbury and the ancient Stoics had a profoundly teleological conception of nature. This is quite explicit in Shaftesbury (I.i.1, p. 167):

We know that every Creature has a private Good and Interest of his own; which Nature has compel’d him to seek by all the Advantages afforded him, within the compass of his Make. We know that there is in reality a right and a wrong State of every Creature; and that his right-one is by Nature forwarded, and by himself

---

5 Citations of this work are to book, part and section number, and to the page numbers in Shaftesbury (Anthony Ashley Cooper, 3rd Earl of Shaftesbury), Characteristics of Men, Manners, Opinions, Times, ed. Lawrence Klein (Cambridge University Press, 2003).

6 See Shaftesbury’s Inquiry Concerning Virtue or Merit, I.i.1, p. 167.

7 See e.g. Shaftesbury’s observation that even “ruffians” have a sense of honour, in the Inquiry Concerning Virtue or Merit, I.ii.4, p. 177.
affectionately sought. There being therefore in every Creature a certain Interest or Good; there must be also a certain END, to which every thing in his Constitution must naturally refer.

Thus, the “constitution” or “make” of each individual creature involves some natural “end” or purpose towards which its nature is directed or oriented. Moreover, according to Shaftesbury, we should not just look to the natural end of the nature of each individual creature; we should look to the natural end of the whole species (I.ii.1, p. 168):

If therefore in the Structure of this or any other Animal, there be any thing which points beyond himself, and by which he is plainly discover’d to have relation to some other Being or Nature besides his own; then will this Animal undoubtedly be esteem’d a Part of some other System. For instance, if an Animal has the Proportions of a Male, it shews he has relation to a Female. … So that the Creatures are both of ’em to be consider’d as Parts of another System: which is that of a particular Race or Species of living Creatures, who have some one common Nature, or are provided for, by some one Order or Constitution of things subsisting together, and co-operating towards their Conservation, and Support.

Thus, Shaftesbury thinks that “there is a System of all Animals; an Animal-Order or Oeconomy, according to which the animal Affairs are regulated and dispos’d”; indeed, he even speculates that the entire universe as a whole may form a single “System” that has some natural end or purpose (I.ii.1, p. 169).

Butler’s Sermons seem to endorse the basic ideas behind Shaftesbury’s teleological conception of nature. Thus, Butler says in the Preface, clearly following Shaftesbury’s idea that the “parts” of a “system” form an overall “economy” or “constitution” (P 14):

Whoever thinks it worth while to consider this matter thoroughly, should begin with stating to himself exactly the idea of a system, economy, or constitution of any particular nature, or particular any thing: and he will, I suppose, find, that it is an one or a whole, made up of several parts; but yet, that the several parts even considered as a whole do not complete the idea, unless in the notion of a whole you include the relations and respects which those parts have to each other. Every work both of nature and of art is a system: and as every particular thing, both natural and artificial, is for
some use or purpose out of and beyond itself, one may add, to what has been already brought into the idea of a system, its conduciveness to this one or more ends.

Butler famously illustrates this conception of “nature” with the example of a watch. To understand a watch properly (or to have the complete “idea” of a watch, as Butler puts it), one must not only know what its parts are, but what their mutual relations are, and how this arrangement of these parts makes the whole watch conducive to the “end” or “purpose” of the watch. In a similar way, Butler promises, his study of human nature will lead to an understanding or “idea” of a human being, and “from this idea itself it will as fully appear that this our nature, i.e. constitution, is adapted to virtue, as from the idea of a watch it appears that its nature, i.e. constitution or system, is adapted to measure time” (P 14).

The same teleological conception of nature is set out at the beginning of Sermon II (II.1):

If the real nature of any creature leads him and is adapted to such and such purposes only, or more than to any other; this is a reason to believe the Author of that nature intended it for those purposes. Thus there is no doubt the eye was intended for us to see with. And the more complex any constitution is, and the greater variety of parts there are which thus tend to some one end, the stronger is the proof that such end was designed.

Admittedly, there is one way in which Butler’s teleological conception clearly differs from Shaftesbury’s. Whereas Shaftesbury is uninhibited about speculating about the natural end or purpose of the whole species, or even of the whole universe, Butler — a vastly more cautious thinker than Shaftesbury in almost every way — avoids committing himself about such large questions, and focuses exclusively on the natural end or purpose of the constitution of the individual human being. Apart from this difference, however, Butler’s teleology seems fundamentally similar to Shaftesbury’s.

It might seem surprising to some readers that this teleological conception of nature
plays such a fundamental role in the works of Shaftesbury and Butler. Teleological conceptions of nature are often thought to belong to a “pre-modern” world view, which — it is often thought — was swept away with the rise of the new science that was typified by Galileo’s *Dialogue Concerning the Two Chief World Systems* (1632). However, this is a very partial view of the period in question. It is true that Galileo’s new science inspired some philosophers — most notably, Hobbes — to turn their back on teleological thinking. But this was certainly not how all philosophers in this period responded. Most thinkers who had a serious interest in biology or medicine would have found it indispensable to appeal to natural ends or purposes in trying to understand the nature of living things. Even in the last decade of the 18th century, Immanuel Kant devoted the entire second part of his *Critique of Judgment* (1790) to an attempt to understand how it can be legitimate for natural scientists to take a teleological approach to understanding the empirical world. It was not the new science of Galileo that finally swept away the appeal to traditional teleology in natural science, but Charles Darwin’s *Origin of Species* (1859) more than two hundred years later.

Admittedly, the mechanistic forms of physics that came to be accepted in the 17th century did indeed avoid the sort of teleological theorizing that had been common in earlier attempts to understand the natural world. However, as recent historians of science have shown, whereas scientists in the first half of the 17th century were optimistic that their mechanistic form of physics would ultimately provide a *complete* explanation of the entire universe, the rise of Newtonian mechanics later in the century actually *discouraged* any such belief in the causal completeness of physics.8 It seemed perfectly reasonable to the best-educated thinkers in the late 17th and early 18th centuries to hypothesize that in addition to the

---

forces that were studied by Newtonian mechanics, there should be other fundamental forces (such as “vital forces”) that were distinctive of living things, and that the best way to understand these forces would be by means of a teleological approach.

2. The interpretation of Butler’s teleology

Clearly, it is crucial for interpreting Butler correctly, then, to understand exactly what this teleological conception of human nature amounts to. How exactly does Butler conceive of what it is for the constitution or nature of something to involve a certain “end” or “purpose”? Different commentators have interpreted Butler’s teleology in different ways. One simple and straightforward interpretation focuses on Butler’s reference to what “the Author of nature intended”; according to this interpretation, for the constitution or nature of an object to involve a certain end or purpose is simply for the creator of that object to have *designed* it to promote that end. This is how Stephen Darwall interprets Butler’s teleology, for example. It is because Darwall interprets Butler’s teleology in this way that he raises the following objection to Butler: “It is difficult to see … how any facts about functional design can establish [Butler’s conclusion], since no normative facts follow from them” (p. 267).

Someone might offer the following reply to Darwall on Butler’s behalf. Since Butler accepts the fundamental articles of traditional Christian belief, he believes that everything in the world is created by an all-knowing, almighty, and morally perfect God. Thus, if God designed us for a certain purpose, it must be right and proper for this purpose to be the supreme purpose of our existence. Thus, given these traditional Christian assumptions, facts

---

about our functional design do imply normative facts.

Darwall might retort that this reply on behalf of Butler suffers from the following problem. Butler rejects a “voluntarist” interpretation of God’s will. If God intends us to lead a certain sort of life, His will is not arbitrary. On the contrary, if He intends us to lead a certain sort of life, He does so precisely because this is the life that there is overriding reason for us to lead. So it is not God’s intending us to lead a virtuous life that makes it the case that there is overriding reason for us to lead a virtuous life. The fact that God intended us to lead virtuous lives is, at most, decisive evidence that there is overriding reason for us to be virtuous. It does not explain why there is such an overriding reason for us to be virtuous.

However, it is not at all clear to me that Butler is trying to give an explanation of why we have an “obligation to the practice of virtue”. That question seems to be the topic for the sort of philosophical speculation that Butler would most likely regard as unnecessary for his ultimate homiletic project. It would be enough for this project if Butler can produce a compelling argument for the conclusion that we have such an “obligation to the practice of virtue”, without also going on to speculate about why this is the case.

Be that as it may, however, it seems to me that from Butler’s point of view, there is a more serious problem with the argument that Darwall ascribes to him. How could Butler think he knows anything about God’s intentions? There are many passages where Butler seems extremely wary about speculating about God’s intentions. For example, he describes the supposition “that the end of divine punishment is no other than that of civil punishment, namely, to prevent future mischief” as a “bold supposition”, “which it would be very presumptuous to assert” (P 29). So what is different about those cases where he is willing to

---

10 For evidence of Butler’s rejection of theological voluntarism, see XIV.14, and The Analogy of Religion, I.v.12n.
make claims about God’s intentions with respect to a “system”, such as the human eye (II.1)? It seems that Butler thought it perfectly *obvious* that a teleological conception of the human eye was correct: the nature of the eye essentially involves a certain end or purpose — namely, for us to see with. Given his traditional Christian assumptions, Butler can then infer from this conception of the eye, as having a nature that essentially involves this end or purpose, to the conclusion that God must have created the eye with this nature, and so must have intended it for this end or purpose.

However, if the claim that the nature of a system involves a certain end or purpose is to *support* the conclusion that God intended the system for that purpose, it must be possible to have *independent* reasons for accepting this teleological claim about this system — that is, reasons for accepting this claim that do not depend on any assumptions about the Creator’s intentions with respect to this system. So the teleological claim that the nature of the system involves that end or purpose surely cannot just *mean* that the system was originally designed for that end or purpose. Indeed, it is clear that when the ancient philosophers made teleological claims of this sort, they were not just making claims about the Creator’s intentions. Many ancient philosophers firmly believed that there are many things the nature of which involves an end or purpose, but did not believe that those things were designed by a Creator for that purpose.11

A different interpretation of Butler’s teleology is suggested by the way in which Shaftesbury expresses his teleological conception: “We know that there is in reality a right and a wrong state of every creature; and that [the creature’s] right one is by nature forwarded” (I.ii.1, p. 167). Moreover, this sort of teleological conception is true, Shaftesbury

---

11 The most striking example of this is Aristotle, whose conception of nature was profoundly teleological, but was not based on any assumption about the Creator’s intentions.
believes, of everything in the universe that has enough of a “constitution” to count as a “system”.

This suggests a quite different interpretation of Butler’s teleology. On this interpretation, a teleological conception of a certain “system” holds that there is a right or correct or proper way for that system to operate in, and it is a fundamental principle governing the behaviour of that system that it is generally disposed to operate in the way that counts as the right or proper way for it to operate in. This way of operating is the “end” or “purpose” towards which the system is oriented or “adapted”.

This sort of teleology can certainly allow that these dispositions that are conducive to the system’s end or purpose may be inhibited or blocked by various interfering factors; these dispositions do not have to be manifested in every possible case. As Butler says, even a watch “is apt to be out of order” (P 14). This does not prevent it from being the case that the watch is generally disposed or “adapted” to tell the time. It is a common occurrence that a thing’s dispositions are inhibited or blocked in this way; that does not prevent the thing from genuinely possessing the dispositions in question.12

This interpretation of Butler’s teleology is in effect closely related to Mark Bedau’s interpretation of what teleological explanations amount to.13 According to Bedau, the defining feature of a teleological explanation is that it seeks to explain a contingent event by showing what is good about that event. On this interpretation, then, the proponent of such a teleological explanation is committed to the view it is a basic feature of the natural system in question that contingent events can occur within that system precisely because it is good for

12 This point has been stressed by much recent work on dispositions. See especially Alexander Bird, “Dispositions and Antidotes”, *Philosophical Quarterly* 48 (1998), 227–34.

them to occur. For example, the plants put out leaves because it is good for them to do so; predators grow sharp teeth because it is good for them to do so. (Some thinkers might even extend this teleological approach beyond the realm of biology, for example suggesting that the rain falls because it is good for it to help the plants to grow.) In general, according to a teleological approach, the goodness of some possible event can make that event actually occur, because the natural system in question has a fundamental tendency to operate in the way that is right or best for it to operate in.

If this is the right way to understand Butler’s teleology, then we can see that Darwall was quite mistaken to claim that “no normative facts follow from” a teleological conception of a system. On the contrary, a teleological conception of a system essentially incorporates a conception of what is the right or proper state for that system — that is, the state that the system ought to be in. In that sense, a teleological conception involves a normative conception. A teleological fact — a fact about what teleological conception is correct — does indeed imply a normative fact.

However, it does not follow that we can only discover the teleological fact on the basis of a prior knowledge of the normative fact. On the contrary, we may know that some teleological conception of a certain system is correct; then an investigation of the dispositions that seem most fundamental to and characteristic of that system may help us to see precisely which of the states that could be intelligibly regarded as the right or proper state for that system to be in is the “end” that those dispositions are conducive to.

To Butler, I propose, it must simply have seemed obvious that some teleological conception of the human mind must be correct. After all, Butler assumes that it will seem obvious to everyone that a teleological conception of the human eye must be correct (II.1); and would it not be extraordinary if the human eye had a natural end or purpose but the
human mind did not? So, an empirical investigation of the most characteristic and fundamental dispositions of the human mind should help us to see which, out of all the many ways of life that could be intelligibly regarded as the right or proper way for a human being to live, is the natural “end” that these fundamental dispositions are conducive to. We may then conclude that it is not just our end or purpose to lead this way of life, but it is also the right and proper way for us to live; it is the way of life that we have overriding reason to lead.

Some philosophers might think that this teleological conception of the human mind is just too antiquated to take seriously. So perhaps Butler’s moral philosophy belongs to those parts of the history of philosophy that the progress of science has rendered utterly obsolete?

We should agree, I think, that Shaftesbury’s teleology, according to which whole species (and perhaps even the whole system of the universe as a whole) has a natural end or purpose, has indeed been rendered obsolete by the advances of contemporary natural science. Evolutionary biologists have incontrovertibly shown that the theory of evolution through natural selection provides vastly more powerful and empirically adequate explanations of biological phenomena than any traditional teleological appeal to a fundamental tendency of living things towards leading the sort of life that it is right and best for them to lead.

However, it is not so clear that the aspect of Butler’s teleology that is most central to the argument of his Sermons — his teleological conception of the individual human mind — has also been shown to be obsolete by contemporary natural science. Indeed, Butler’s teleological conception of the human mind could be defended on the basis of a philosophy of mind that accepts some version of the slogan that “the intentional is normative”. For example, a Davidsonian philosophy of mind would insist that we must have some tendency to believe the truth, and to love the good, if we are to be correctly interpretable as having the
attitudes of belief or love at all.\textsuperscript{14} According to a philosophy of mind of this sort, it is essential to the various types of mental state that are characteristic of the human mind that there is a correct or proper role that these mental states should play in human thinking and reasoning, and in every human being, these mental states must have at least some disposition to play this correct or proper role, if the human being is to be capable of mental states of that sort at all.\textsuperscript{15} In effect, this philosophy of mind adopts a teleological conception of the mind, of the same general kind as I have ascribed to Butler. So the proponent of this sort of philosophy of mind should be able to welcome the naturalist project that Butler pursues in the \textit{Sermons}.

3. Butler’s conception of reasons for action

In this section, I shall highlight a further feature of Butler’s project, which sharply distinguishes Butler from that many later philosophers of the modern era. This feature concerns Butler’s conception of \textit{reasons for action}.

As we have seen, Butler’s aims to base his argument for the conclusion that we have an overriding reason (or “obligation”) to be virtuous on an account of \textit{human nature}. So, according to Butler, \textit{all human beings} have an overriding reason to be virtuous. This marks an important difference between Butler’s approach and the approach of many other modern philosophers. On the one hand, Butler’s approach differs from that of philosophers of a broadly Humean persuasion, such as Philippa Foot (at least in the 1970s) and Bernard

\textsuperscript{14}As Donald Davidson put it, in interpreting someone we “will try for a theory that finds him consistent, a believer of truths, and a lover of the good”; see his “Mental Events”, \textit{Essays on Actions and Events} (Oxford University Press, 1980), p. 222.

\textsuperscript{15}For a defence of this sort of philosophy of mind, see my forthcoming paper “The Normativity of the Intentional”, in Brian McLaughlin and Ansgar Beckermann, eds. \textit{The Oxford Handbook of the Philosophy of Mind} (Oxford University Press, forthcoming).
Williams, since these philosophers hold that only human beings who have certain desires or interests (or who have appropriate elements in their “subjective motivational set”) have a reason to be virtuous. At the same time, Butler’s approach also differs from that of philosophers of a broadly Kantian persuasion, such as Christine Korsgaard, who base their conclusion that we all have overriding reason to be virtuous not on an account of human nature, but rather on an account of the necessary structure of rational agency as such.

Underlying this difference between Butler, on the one hand, and the Humeans and Kantians, on the other hand, is the fact that the Humeans and Kantians both seem to hold a purely formal or procedural conception of reasons for action. To the extent that these philosophers are willing to make sense of the notion at all, they conceive of what there is “overriding reason for one to do” in terms of what it is formally or procedurally rational for one to do. First, these philosophers start with a notion of what it is for a process of practical reasoning to count as procedurally rational. Then, they propose, there is “overriding reason” for one to perform an action just in case, if one were adequately informed of the relevant non-

---

16 See Philippa Foot, “Morality as a System of Hypothetical Imperatives”, in her *Virtues and Vices* (Oxford: Blackwell, 1978), and Bernard Williams, “Internal and External Reasons”, in his *Moral Luck* (Cambridge University Press, 1982). For the Humean antecedents of this position, see David Hume, *A Treatise of Human Nature*, II.iii.3. In her later work, Philippa Foot argues for a position that is closer to Butler’s, in that it sees our reason to comply with moral requirements as ultimately grounded in human nature; see her *Natural Goodness* (Oxford University Press, 2001). For another approach that also grounds our reason to comply with moral requirements in human nature, see Paul Bloomfield, *Moral Reality* (Oxford University Press, 2001).

17 See Christine Korsgaard, *The Sources of Normativity* (Cambridge University Press, 1996). Even though Korsgaard appeals to our “practical identity” as human beings, she argues that any rational agent who is capable of reflective choice at all is committed to recognizing this practical identity; so she is not basing a defence of virtue on an empirical conception of human nature in the way that Butler is doing.

normative facts about one’s situation, and went through a process of procedurally rational practical reasoning, one would choose to perform that action. Of course, the Humeans and the Kantians have strikingly different conceptions of what it is for a process of practical reasoning to count as “procedurally rational”. But there are certain fundamental similarities: for example, both Humeans and Kantians believe that the conditions of procedurally rational practical reasoning have a similar status to the laws of logic.\textsuperscript{19} I shall refer to these philosophers’ conceptions of reasons for action as \textit{procedural conceptions}.\textsuperscript{20}

On the face of it, however, the Kantian claim — that, for absolutely all well-informed rational agents, it is procedurally irrational to violate a moral requirement, in essentially the same way as it is procedurally irrational to violate the laws of logic — is an awfully strong claim. Offhand, it seems possible for an agent to violate a moral requirement even if his reasoning is logically quite coherent and free from any error or ignorance about the relevant non-normative facts. For example, we could imagine a brilliantly successful criminal. Suppose that this criminal is a genius at \textit{a priori} reasoning — at mathematics, logic, decision theory, and so on — but he does not accept that moral requirements provide him with any reason to act accordingly, and has committed appalling crimes without compunction or remorse. Is it really necessary that this criminal’s practical reasoning is either in some way \textit{procedurally irrational}, or else \textit{misinformed} or \textit{ignorant} about some relevant non-normative
fact? The claim that there must be some such procedural irrationality or non-normative error or ignorance in his reasoning surely has a heavy burden of proof to bear. Many Kantians are willing to try to shoulder this burden of proof. Prima facie, however, this burden of proof gives us a reason to try to find an acceptable alternative to this Kantian approach.

In general, the procedural conception of reasons for action seems to make it at least prima facie implausible that moral requirements must provide a consistent egoist with strong or weighty reasons to act accordingly. As the Humeans claim, with considerable prima facie plausibility, even if a criminal or an egoist is procedurally rational and ideally well informed about the relevant non-normative facts, he could still lack any motivations that would lead him to choose to comply with moral requirements. So, the Humeans conclude, moral requirements would not provide such an egoist with any reasons to act at all.

However, this Humean view also seems open to prima facie serious objections. James Doyle puts it well:

> The point is not just that [on this Humean view] there will be nothing we can say by way of rational persuasion on behalf of morality to someone, such as the egoist, who just happens to lack the relevant motivation — although this is true. … The real problem with such a Humean view is that we will not even have anything to say to each other about what mistake, exactly, the egoist is making.

On this Humean view, the egoist is making no mistake at all: he is quite right to deny that moral requirements provide him with any reasons whatsoever. Indeed, you would be right to deny that moral requirements provide you with any reasons, if you too came to lack the

---

21 Besides Korsgaard, the most notable Kantian who has attempted to shoulder this burden of proof in recent years is Thomas Nagel, in The Possibility of Altruism (Princeton University Press, 1970).

relevant motivations. But this is surely not an attractive way to conceive of morality. So if we embrace the procedural conception of reasons for action, it will be hard for us to give a satisfactory account of our reasons to comply with moral requirements: either we will have to shoulder the heavy burden of proof that the Kantians must bear, or we will be forced into the unattractive conclusion of the Humeans.

Butler does not accept the procedural conception of reasons for action. In his view, the egoist, in acting viciously, is acting as there is overriding reason for him not to act. But Butler never claims that the egoist is *procedurally irrational*. Unlike Samuel Clarke, Butler never claims that anyone who acts viciously is “guilty of the very same unreasonableness and contradiction in one case; as he that in another case should affirm one number or quantity to be equal to another, and yet that other at the same time not to be equal to the first.”\(^{23}\) For all that Butler says, there need be no procedural irrationality, and no logical incoherence, in the egoist’s thinking at all.

Butler also never claims that the egoist is necessarily misinformed or ignorant about any *non-normative* fact. Of course, if the egoist believes that moral requirements do not provide him with any reasons to act accordingly, then according to Butler, the egoist believes something *false*. But this is a false normative belief, not about a false belief about any non-normative fact. Butler does not claim that this false normative belief must be explained either by procedural irrationality or by error or ignorance about any non-normative fact.

\(^{23}\) See D. D. Raphael, ed., *The British Moralists* (Oxford University Press, 1969), §232. (Strictly speaking, Clarke is only referring to *injustice* here, not to vice in general, although elsewhere he claims that there is a similar “absurdity and inconsistency” in other kinds of vice as well.) It is true that Butler concedes that Clarke’s method of “inquiring into the abstract relations of things” is just as valid as his own method of inquiring into human nature (P 12). However, there is no reason to interpret Butler’s endorsement of Clarke’s method as an endorsement of Clarke’s view that anyone who acts viciously is guilty of “absurdity and inconsistency”.
Moreover, Butler may have a positive reason for being sceptical of any such procedural conception of reasons. Butler seems to be generally quite sceptical of attempts to give reductive definitions. Arguably, this is part of what he meant when he famously said, “Everything is what it is, and not another thing” (P 39). Thus, Butler would be equally sceptical of attempts to reduce the notion of “what there is overriding reason for one to do” to the notion of “what one would choose to do if one were procedurally rational and adequately informed about the non-normative facts”. Butler could probably endorse the following elucidation of the notion of “what there is overriding reason for one to do”. The judgment that there is “overriding reason” for one to do something expresses a conclusion of practical reasoning. So it would be weak-willed or akratic for one to form the judgment that there is “overriding reason” for one to do a certain thing, but then willingly to fail to do it. But Butler would not accept any attempt to reduce the notion of an “overriding reason” to the notion of “procedural rationality”. An overriding reason for action is just an overriding reason for action; it is what it is, and not another thing.

Of course, Butler can still claim that one way in which one might fail to recognize what there is overriding reason to do is by being procedurally irrational. But if he rejects the procedural conception of reasons for action, then he must regard it as possible, at least in principle, that even if one is procedurally rational, and ideally well informed about all relevant non-normative facts, one could still fail to recognize what there is overriding reason to do. If such failures are possible, this must be because our most basic normative beliefs arise, not just from procedurally rational reasoning from non-normative premisses, but from a specific faculty that can malfunction even if one is procedurally quite rational.

Butler certainly believes in such a specific faculty. He calls this faculty ‘conscience’ (I.8), although sometimes he also uses other names, such as ‘reflex approbation or
disapprobation’ or ‘reflection’ (P 26). He acknowledges that there seems to be “some small diversity amongst mankind” with respect to the deliverances of conscience (II.1), which seems to imply that our conscience is fallible.\footnote{Thus, Butler is not committed to denying that conscience is fallible — contrary to what G. E. M. Anscombe says in “Modern Moral Philosophy”, \textit{Philosophy} 33 (1958): 1–19. However, Butler does not develop this point. He clearly thinks that the main danger to our appreciation of moral and normative truths is \textit{self-deception}, which he discusses at length in Sermons VII and X.} Someone whose conscience fails to inform them of a normative truth might have a corrupted or defective conscience, but he need not count either as procedurally irrational or as misinformed about any non-normative truth.

In short, Butler’s argument for “our obligation to the practice of virtue” differs from that of most later modern philosophers, in the following ways. Unlike the Humeans, Butler argues for the conclusion that moral requirements really do provide the egoist with overriding reasons to act accordingly; unlike the Kantians, Butler bases his argument for this conclusion in an account of human nature, not in an account of the necessary structure of rational agency as such; and unlike most later moral philosophers, including both the Humeans and the Kantians, he does not accept the procedural conception of reasons.

4. Butler’s rejection of eudaimonism

Even though Butler’s approach differs in this way from the approach of most later modern philosophers, there is also one crucial way in which his approach also differs from that of the “ancient moralists” whom he claims to be following. Unlike them, Butler \textit{does not accept eudaimonism}. As I shall understand it, eudaimonism is a view about reasons for action, according to which it is universal principle, applying to all agents and all actions, that an agent has an overriding reason to perform an action if and only if that action \textit{promotes the}
agent’s happiness more than any available alternative action.

It is important not to misread this formulation of eudaimonism. According to this formulation, eudaimonism does not make any claim about why one has an overriding reason to perform these actions, or about what is the proper motive for performing these actions. So, in particular, eudaimonism does not claim that the only reason that there is for performing these actions is that these actions will best promote one’s own happiness, or that the proper motive for performing them is the desire to promote one’s happiness. On the contrary, it is perfectly compatible with eudaimonism to claim that the fundamental reason for performing many of these actions, and the reason that ought to motivate one to perform them, is just that these actions are intrinsically fine or admirable. Eudaimonism only claims that whenever one has overriding reason to perform an action, that action will also promote one’s happiness more than any available alternative; and conversely, whenever an action will promote one’s happiness more than any available alternative, one has an overriding reason to perform it. Eudaimonism claims that overriding reasons for action perfectly coincide with the demands of one’s own happiness: it does not claim that these reasons for action all arise from the demands of one’s own happiness.

Even if we understand eudaimonism in this cautious way, however, Butler does not rest his argument for our “obligation to the practice of virtue” on this eudaimonist principle. Instead, he bases his argument on his own principle of the natural supremacy of conscience — that is, the principle that it is an essential part of human nature that our conscience should be supreme.25

This point emerges most clearly in a passage where Butler criticizes what he calls “a

25 This principle appears to imply that it is an essential part of human nature to have a conscience. Any member of our species (e.g. an infant) who lacked a conscience would not be a full-blown instance of “human nature” in the relevant sense.
There are other passages that seem to recognize such “exceptions”. E.g.: “Self-love then, though confined to the interest of the present world, does in general perfectly coincide with virtue; and leads us to one and the same course of life. But, whatever exceptions there are to this, which are much fewer than they are commonly thought …” (III.8).

[Lord Shaftesbury] has shewn beyond all contradiction, that virtue is naturally the interest or happiness, and vice the misery, of such a creature as man, placed in the circumstances which we are in this world. But suppose there are particular exceptions; a case which this author was unwilling to put, and yet surely it is to be put: or suppose a case which he has put and determined, that of a sceptic not convinced of this happy tendency of virtue, or being of a contrary opinion. His determination is, that it would be without remedy. One may say more explicitly, that leaving out the authority of reflex approbation or disapprobation, such an one would be under an obligation to act viciously; since interest, one’s own happiness, is a manifest obligation, and there is not supposed to be any other obligation in the case.

Here, Butler seems to think that we must allow, at least for the sake of argument, that there are “exceptions” to the general rule that virtue and self-interest coincide; and he seems to want to develop an argument for the conclusion that we have “an obligation to the practice of virtue” that could be accepted by “a sceptic not convinced of this happy tendency of virtue, or being of a contrary opinion”. Such a sceptic presumably rejects the traditional Christian doctrine that we will all receive rewards or punishments in an after-life that will ensure the perfect coincidence of virtue and happiness. So Butler’s argument for his conclusion is designed to be acceptable to someone who rejects the Christian doctrine of an after-life.

By an “exception” here Butler clearly means an exception to Shaftesbury’s general conclusion that “virtue is naturally the interest or happiness, and vice the misery, of such a creature as man, placed in the circumstances which we are in this world”. Now, in his *Inquiry concerning Virtue or Merit*, Shaftesbury is primarily concerned, not with a comparison between particular actions, but with a comparison between overall ways of life. Thus, Shaftesbury’s conclusion is that virtuous ways of life are, in general, happier than vicious

---

26 There are other passages that seem to recognize such “exceptions”. E.g.: “Self-love then, though confined to the interest of the present world, does in general perfectly coincide with virtue; and leads us to one and the same course of life. But, whatever exceptions there are to this, which are much fewer than they are commonly thought …” (III.8).
ways of life. So an “exception” to Shaftesbury’s conclusion would be a case in which someone will be happier on the whole if he leads a certain vicious way of life than if he leads a virtuous way of life. Presumably, if there is a vicious way of life that will make this person happier than any virtuous way of life, then there will be cases in which some particular vicious action will make the person happier than any available virtuous alternative action. Butler’s argument is designed to be compatible with the existence of such “exceptions”. So his argument must be compatible with the existence of cases in which one has an overriding reason to act virtuously but the virtuous action does not promote one’s happiness more than any alternative. Thus, his argument must be compatible with the falsity of eudaimonism.

In this passage, then, Butler is implying that once we take account of the natural supremacy of conscience, we can solve the problem that when faced with an “exception” to the general coincidence of virtue and self-interest, one would be “under an obligation to act viciously”. Immediately after this passage, Butler considers an objection to this view, and then offers a rather unexpected reply to that objection (P 26):

‘But does it much mend the matter, to take in that natural authority of reflection? There indeed would be an obligation to virtue; but would not the obligation from supposed interest on the side of vice remain?’ If it should, yet to be under two contrary obligations, i.e. under none at all, would not be exactly the same, as to be under a formal obligation to be vicious, or to be in circumstances in which the constitution of man’s nature plainly required that vice should be preferred. But the obligation on the side of interest really does not remain. For the natural authority of the principle of reflection is an obligation the most near and intimate, the most certain and known: whereas the contrary obligation can at the utmost appear no more than probable; since no man can be certain in any circumstances that vice is his interest in the present world, much less can he be certain against another: and thus the certain obligation would entirely supersede and destroy the uncertain one; which yet would have been of real force without the former.

Here, Butler concedes, at least for the sake of argument, that there may be cases in which it is “probable” that it is in one’s interest to be vicious. Nonetheless, even in those cases, it would
still be “certain” that one has an obligation to be virtuous. Hence, the certain obligation to be virtuous completely trumps and removes what would otherwise have been an uncertain and merely probable obligation to be vicious.

Butler concludes this discussion of Shaftesbury’s views as follows (P 27):

In truth, the taking in this consideration totally changes the whole state of the case; and shews, what this author does not seem to have been aware of, that the greatest degree of scepticism which he thought possible will still leave men under the strictest moral obligations, whatever their opinion be concerning the happiness of virtue.

Thus, Butler clearly intends his argument to be completely independent of the eudaimonist view that overriding reasons for action always coincide with the demands of one’s own happiness. Even though the conclusion of his argument is that every human being has an overriding reason to act virtuously at all times, his argument is designed to be compatible with the existence of cases (“exceptions”) in which it is “probable” that one will promote one’s own happiness more effectively by being vicious than by being virtuous. Butler’s only concession to eudaimonism is to accept that if one has an overriding reason to pursue a certain course of action, then it cannot be certain that refraining from that course of action will promote one’s happiness more. (We shall inquire later on exactly why Butler makes this limited concession to eudaimonism.)

As we shall see later, there is another feature of Butler’s Sermons that also reveals Butler’s refusal to accept eudaimonism. This feature emerges in the fact that he distinguishes sharply between conscience and self-love. It is conscience that makes us aware of, and inclines us to pursue, the way of life that we have overriding reason to lead, while it is self-love that makes us aware of, and inclines us to pursue, our own self-interest or happiness. Butler resists any attempt to identify conscience with self-love, or to view either of these two “inward principles” as merely a superfluous adjunct to the other. Since these two faculties are
independent in this way, our judgments about what we have overriding reasons to do arise independently of our judgments about what will promote our own happiness; so there is no reason at this stage in the argument to assume that the two sorts of judgments will universally and perfectly coincide.

5. The supreme authority of conscience

As I have argued above, Butler’s project involves arguing for a certain teleological conception of human nature. According to such a teleological conception, it is part of human nature that there is a certain sort of life that is the right or correct or proper life for any human being to lead. To deviate from this sort of life, Butler claims, would be “disproportionate to the nature of man”, and so “in the strictest and most proper sense unnatural” (II.10). Butler appears to assume that it is a basic truth about reasons for action that the life that is, in this sense, the correct or proper life for human beings to lead is also the life that there is overriding reason for human beings to lead. What Butler aims to show is that this sort of life essentially involves living virtuously and complying with moral requirements. In fact, he aims to show that this sort of life is not only a virtuous life, but also a life that involves the effective pursuit of the agent’s own happiness.

It is part of any teleological conception of human nature (according to my interpretation of Butler’s teleology) that the human mind is a “system” in which various elements are structured in relation to each other in such a way that the whole mind is generally disposed or “adapted” to leading the kind of life that is the correct or proper life for it to lead. This is why an empirical investigation of our mental dispositions can help us to see what is the correct or proper life for us to lead.
Thus, Butler has to argue that in various ways, many of our mental dispositions are conducive to virtue, and many of these dispositions are also conducive to the effective pursuit of happiness. Butler gives an initial summary of this argument in Sermon I (I.15):

The nature of man considered in his single capacity, and with respect only to the present world, is adapted and leads him to attain the greatest happiness he can for himself in the present world. The nature of man considered in his public or social capacity leads him to a right behaviour in society, to that course of life which we call virtue. Men follow or obey their nature in both these capacities and respects to a certain degree, but not entirely: their actions do not come up to the whole of what their nature leads them to in either of these capacities or respects …

This point is also argued at greater length throughout the Sermons. For example, in Sermon IV, he focuses on our disposition to talkativeness — roughly, our tendency to like the sound of our own voices — and argues that our delight in idle chatter helps us to cement the social bonds between us. In Sermons V and VI, he focuses on compassion, our tendency to feel the pain of others, and argues that this helps us to be charitable and inhibits us from cruelty. In Sermons VIII and IX, he focuses on anger and resentment: here, he argues that the natural end or purpose of our tendency to “sudden anger” is “self-defence” (VIII.6); while “deliberate” or “settled resentment … is to be considered as a weapon, put into our hands by nature, against injury, injustice and cruelty” (VIII.8). In Sermon XI, he considers self-love, and argues that self-love is not essentially in tension with virtue, and can even play a role in supporting virtue. Finally, in Sermons XII-XIV, he considers mental dispositions that have a more obvious role in supporting virtue — namely, the “principle of benevolence” (in Sermon XII), and our disposition towards religious feeling or the “love of God” (in Sermons XIII and XIV). In general, he says, “every one of our passions and affections hath its natural stint and bound” (XI.9); within these natural bounds, none of these passions conflicts with virtue, and most passions help us, at least indirectly, either to be virtuous or to pursue our own happiness
However, the most distinctive element in Butler’s argument concerns the special role of conscience, which is his focus in Sermons II and III. It seems to be part of Butler’s teleological conception of human nature that the right or correct or proper sort of life for human beings to lead must involve all the various elements of human nature functioning in a certain way — as we might put it, it involves all these elements functioning “properly”. The elements of human nature that Butler focuses on are what he calls the “internal principles” of the human mind, of which he gives a brief survey in Sermon I (I.5-8). These “internal principles” include: the principles of benevolence and self-love (I.6); our various particular passions (I.7); and conscience (I.8).

In Sermon I, Butler describes a preliminary description of how conscience actually operates (I.8):

There is a principle of reflection in men, by which they distinguish between, approve and disapprove their own actions. We are plainly constituted such sort of creatures as to reflect upon our own nature. The mind can take a view of what passes within itself, its propensions, aversions, passions, affections, as respecting such objects, and in such degrees; and of the several actions consequent thereupon. In this survey it approves of one, disapproves of another, and towards a third is affected in neither of these ways, but is quite indifferent. This principle in man, by which he approves or disapproves his heart, temper, and actions, is conscience …. And that this faculty tends to restrain men from doing mischief to each other, and leads them to do good, is too manifest to need being insisted upon. … It is needless to compare the respect it has to private good, with the respect it has to public; since it plainly tends as much to the latter as to the former, and is commonly thought to tend chiefly to the latter.

There is a second description of the operations of conscience in Sermon II (II.8):

But there is a superior principle of reflection or conscience in every man, which distinguishes between the internal principles of his heart, as well as his external actions: which passes judgment upon himself and them; pronounces determinately some actions to be in themselves just, right, good; others to be in themselves evil, wrong, unjust: which, without being consulted, without being advised with,
magisterially exerts itself, and approves or condemns him the doer of them accordingly….

Thus, Butler thinks of conscience as a faculty that reviews both our inner mental states (such as our feelings and intentions) and our external actions. Its deliverances are described sometimes as states of “approval” or “disapproval”, and sometimes as “judgments”; it is also credited with the power to “restrain” us from some actions, and to “lead” us to do other things instead. Its general tendency is to approve of courses of action that are favourable either to the good of the community or to the good of the agent (or to both). The crucial point for our purposes is that even though Butler does not analyse the operations of conscience in full detail, we can be confident that he believes that when our conscience is functioning properly, it will approve only of “external actions” and of “internal principles of the heart” that are compatible with virtue.

Butler’s crucial move comes in Sermon II, where he argues that when all these elements in our nature are functioning properly, they will form a hierarchy. First, our particular passions will be directed or regulated by self-love (II.10-11). That is, whenever there is a conflict between self-love and any particular passion, then if the elements of our nature are functioning properly, we will act in accordance with self-love, and not in accordance with the passion.

Second, when the elements of our nature are functioning properly, both the particular passions and self-love will be directed or regulated by conscience (II.12-17). That is, when these elements all function properly, we will always act as our conscience directs us to act,
and we will presumably also try to cultivate the “internal principles of our heart” as conscience directs us to. As he puts it, self-love has greater authority than the particular passions, and conscience has even greater authority than self-love — even if the particular passions or self-love may on many occasions have greater motivational strength (II.14). As he also puts it, self-love is a superior principle compared to the particular passions, and conscience is a superior principle compared to both the particular passions and self-love (II.11). As he says, conscience “was placed within us to be our proper governor, to direct and regulate” all the other “principles, passions and motives of action” (II.15).

Butler argues for this point by claiming that conscience “is to be considered … as from its very nature claiming superiority: insomuch that you cannot form a notion of this faculty, conscience, without taking in judgment, direction, superintendency” (II.14). This argument seems to start from identifying certain features of conscience. One feature is conscience’s “superintendency”, which seems to consist in the way in which conscience considers and oversees all the other internal principles of the mind. Another feature is its capacity for “judgment”, which may consist in the fact that it arrives at all-things-considered judgments about what is right or wrong, good or bad, aiming to taking all relevant considerations into account. The final feature is its capacity for “direction”, which may consist in the fact that we have a fundamental disposition to be moved to action by the directions of our conscience.

It is not completely clear why Butler says that a faculty that has these features “claims superiority from its very nature”. But in the light of the teleological reflections that are so prominent at the beginning of Sermon II, it seems plausible to read this as the claim that these features of conscience — “judgment, direction, superintendency” — make it clear that the natural end or purpose of the faculty of conscience is precisely to serve as our “proper
Darwall (The British Moralists and the Internal ‘Ought’, pp. 256–61) argues that there is an altogether different line of thought in II.16–17, which he interprets as part of Butler’s argument for “Kant’s reciprocity thesis”. This interpretation seems entirely misguided to me. Prima facie, the passage in question makes a fairly weak point: it would be absurd to claim “that there was no distinction to be made between one inward principle and another, but only that of strength” (II.16) — that is, there must be some distinctions between “superior” and “inferior principles”. This point by itself does not tell us that conscience is the supreme principle, nor that conscience approves only of actions that are compatible with virtue (it provides some small degree of support to the conclusion that conscience is the supreme principle, because one way in which conscience might not be supreme is if absolutely no distinctions could be drawn between superior and inferior principles at all).

Thus, Butler has argued that when the elements of our nature are all functioning properly, our conscience will only approve of external actions and internal mental states that are compatible with virtue, and we will only act as our conscience approves (and we will also try to cultivate our internal mental dispositions as conscience directs us to). So, when the elements of our nature are functioning properly, we will always act virtuously, and we will also cultivate virtuous internal mental dispositions in ourselves. Given the background assumption that the life that we have overriding reason to lead is the life in which all these elements of our nature are functioning properly, Butler’s conclusion follows: we have overriding reason to be virtuous.

6. Butler’s claims about the harmony of virtue and self-interest

Now that we have a clearer conception of Butler’s overall project, we can return to the controversial question of how exactly Butler understands the relations between virtue and self-interest.

I have already discussed the passages where Butler seems clearly to reject eudaimonism, such as his discussion of Shaftesbury in the Preface (P 26). But there are also

28 Darwall (The British Moralists and the Internal ‘Ought’, pp. 256–61) argues that there is an altogether different line of thought in II.16–17, which he interprets as part of Butler’s argument for “Kant’s reciprocity thesis”. This interpretation seems entirely misguided to me. Prima facie, the passage in question makes a fairly weak point: it would be absurd to claim “that there was no distinction to be made between one inward principle and another, but only that of strength” (II.16) — that is, there must be some distinctions between “superior” and “inferior principles”. This point by itself does not tell us that conscience is the supreme principle, nor that conscience approves only of actions that are compatible with virtue (it provides some small degree of support to the conclusion that conscience is the supreme principle, because one way in which conscience might not be supreme is if absolutely no distinctions could be drawn between superior and inferior principles at all).
some passages where Butler seems to come much closer to the eudaimonist position. The most notorious of these passages is the following (XI.20):

And to all these things may be added, … there can no access be had to the understanding, but by convincing men, that the course of life we would persuade them to is not contrary to their interest. It may be allowed, without any prejudice to the cause of virtue and religion, that our ideas of happiness and misery are of all our ideas the nearest and most important to us; that they will, nay, if you please, that they ought to prevail over those of order, and beauty, and harmony, and proportion, if there should ever be, as it is impossible there ever should be, any inconsistence between them: though these last too, as expressing the fitness of actions, are real as truth itself. Let it be allowed, though virtue or moral rectitude does indeed consist in affection to and pursuit of what is right and good, as such; yet, that when we sit down in a cool hour, we can neither justify to ourselves this or any other pursuit, till we are convinced that it will be for our happiness, or at least not contrary to it.

Here, Butler asserts that it is “impossible there ever should be … any inconsistence between” (i) “our ideas of happiness and misery” and (ii) our ideas of “order, and beauty, and harmony, and proportion,” which express “the fitness of actions”. He also implies that “it may be allowed, without any prejudice to the cause of virtue” that we cannot “justify” an “affection to and pursuit of what is right and good as such” unless “we are convinced that it will be for our happiness, or at least not contrary to it”. It is not surprising that this passage has convinced many scholars that Butler is a eudaimonist after all.

Thus, we must find an interpretation of Butler’s language that enables us to reconcile all of the following four propositions:

---

29 It seems to me that the only natural way to read this clause is an assertion, parenthetically inserted inside a non-asserted sentence that is within the scope of ‘It may be allowed, without any prejudice to the cause of virtue and religion, that …’.

30 For example, Butler is interpreted as a eudaimonist by H. A. Prichard, in Moral Obligation (Oxford: Clarendon Press, 1949), pp. 96–97. Henry Sidgwick interprets Butler as viewing self-love as co-ordinate with conscience, so that an act can count as reasonable only if it is approved by both self-love and conscience, in his Outlines of the History of Ethics, 3rd edition (London: Macmillan, 1892), p. 196.
1. Butler’s ultimate conclusion, that all human beings have an “obligation”, or overriding reason, to be virtuous;
2. The proposition (which Butler does not assert, but seems to regard as compatible with his argument) that no “pursuit” is justified if it is “contrary” to our happiness;
3. The proposition (which Butler also regards as quite compatible with his argument) that there are “exceptions” to the coincidence of virtue and happiness;
4. The proposition (which Butler asserts) that it is impossible for there to be any “inconsistency” between happiness and virtue.

Finally, we also need to understand how it can be that here Butler is asserting that it is “impossible” for there to be any “inconsistency” between happiness and virtue, whereas in the Preface he seems to claim that his argument could be accepted by a “sceptic not convinced of this happy tendency of virtue” (P 26).

As we saw in discussing the Preface (P 26), Butler’s only clear concession to eudaimonism was to insist that we can never be certain that it is in our interest to be vicious, or to fail to be virtuous. It is possible to read these four troublesome propositions in the light of this claim. When Butler speaks of a pursuit’s being “contrary to” our happiness, he may be using these terms in a special sense. Specifically, for a pursuit to be “contrary to” our happiness in this special sense would be for that pursuit to be certain to make us less happy than some available alternative. So, given that he insists that it is never certain that virtue will make us less happy than vice, he can conclude that virtue is never in this sense “contrary to our happiness”. Moreover, when he speaks of an “inconsistency” between happiness and virtue, he may mean a situation in which it is certain that being vicious will promote one’s happiness more than being virtuous. Then the claim that there cannot be any “inconsistency” between virtue and happiness will be compatible with the idea that there could be “exceptions” to the coincidence of virtue and self-interest, if these “exceptions” have the feature that even if it is possible to be in such an “exceptional” case, one can never be certain that one is in such a case.
Finally, when Butler says that we must consider the case of a “sceptic not convinced of this happy tendency of virtue, or being of a contrary opinion”, it may be that his aim is only to produce an argument that is acceptable to someone who thinks that there are “exceptional” cases where it is probable that the demands of virtue and of happiness diverge, not to produce an argument that is acceptable to someone who thinks that there are cases in which it is certain that the demands of virtue and happiness diverge. Butler’s aim may only be to produce an argument that is acceptable to this more modest sort of “sceptic”.

This reading will also help us to explain why Butler’s claims about the harmony of conscience and self-love do not trivialize his claim that conscience has greater authority than self-love. The certain counsels of self-love will never conflict with the directions of conscience. However, there may be some cases in which self-love gives no certain counsels at all, but only uncertain counsels — possibly including uncertain counsels that conflict with the directions of conscience; in these cases, the supremacy of conscience implies we should always follow the certain directions of conscience. However, this reading still leaves us with two exegetical problems. First, why does Butler think he needs to argue for the impossibility of any “inconsistency” between virtue and happiness here? Second, how exactly does Butler think he can argue for this?

Butler never explicitly says why he thinks that he has to argue for the impossibility of any “inconsistency” between virtue and self-interest. But it is relatively easy to see what his main reason for wanting to argue for this must be.

Butler argues that self-love is a superior principle when compared to the “particular passions”: as he says, “the passions … may be contradicted without violating [our] nature, but [self-love] cannot” (II.10). According to the interpretation that I have proposed, what this means is the following: if all the elements of our nature are functioning properly, then
whenever there is a clear conflict between a particular passion and self-love, we will act in accordance with self-love rather than in accordance with the particular passion. This is why he says, “interest, one’s own happiness, is a manifest obligation” (P 26). But Butler also wants to claim that when all the elements of our nature are functioning properly, we will always act virtuously. So, to avoid self-contradiction, he will have to rule out the possibility of cases in which all the elements of human nature are functioning properly, and self-love is in conflict with some particular passion because self-love clearly motivates us to act viciously, while the particular passion motivates us to act virtuously.

He cannot rule out such cases by arguing that when the elements of our nature are functioning properly, the particular passions never motivate us to act virtuously, since as he stresses, many of our particular passions lead us, at least indirectly, to act virtuously. For example, “desire of the esteem of others … naturally lead[s] us to regulate our behaviour in such a manner as will be of service to our fellow creatures” (I.7). Nor could he rule out these cases by arguing that when all these elements of our nature are functioning properly, self-love never conflicts with the particular passions: making that argument would risk draining the claim that self-love is a superior principle of all content whatsoever. So the only way in which he can rule out these cases is by arguing that when all these elements of our nature are functioning properly, self-love never clearly motivates us to act viciously.

In addition, Butler wants to argue, on the basis of his empirical survey of the “internal principles of the human heart”, that the life that one has overriding reason to lead will not just be a virtuous life; it will also be a life that is reasonably effective at achieving of one’s own happiness (I.15). Since he assumes that one is guaranteed to lead this sort of life when all the elements in one’s nature are functioning properly, he will also have to claim that when all these elements are functioning properly, one will be leading a life that is reasonably effective
at achieving one’s own happiness in this life. So it seems that he must argue that a whole life that involves all these elements functioning properly will never be certain to make one less happy than any alternative way of life.

Thus, there are really two claims that Butler has reason to make about the relation between virtue and happiness. The first claim concerns particular external actions, but is restricted to cases in which all the elements of one’s nature are functioning properly: here the claim is that within this restricted range of cases, a vicious action is never certain to make one happier than all of the virtuous alternatives. The second claim concerns a comparison between the whole way of life that involves these elements all functioning properly (including the overall “temper” or pattern of mental dispositions that is characteristic of that way of life), and alternative ways of life in which some of these elements are not functioning properly: here the claim is that the latter ways of life are never certain to make one happier than the former. Moreover, since he is aiming to produce an argument that is acceptable to a “sceptic” who does not accept the traditional Christian doctrine of an after-life, he must argue that a virtuous life is never certain to make one less happy than a vicious life in this world.

Thus, we can sum up Butler’s claims about the precise sort of harmony that exists between virtue and self-interest as follows. There may be “exceptions” — cases where someone leads a vicious way of life and turns out to be happier than he would have been had he led a virtuous way of life — but one can never be certain that one is in such an exceptional case oneself. Thus no one can ever be under a certain obligation of self-love to lead a vicious way of life. By contrast, because of the natural supremacy of conscience, one is always under a certain obligation to lead a virtuous way of life. Once one is leading a virtuous life, and all the “inward principles of one’s heart” are functioning properly (so that all these inward principles remain within their “natural stint and bound”), then one’s
dispositions will be such that, when one is faced with the choice between particular alternative actions, it will never be certain that a vicious action will make one happier than all the available virtuous actions.

Even though Butler makes these claims about the harmony of virtue and happiness, he is not claiming that one’s only reason to be virtuous is simply to promote one’s own happiness. On the contrary, one’s reason to be virtuous is simply that a virtuous life is the right or proper life for us to lead — as is shown by the fact that when all the elements of our nature are functioning properly, one’s conscience will always direct one to act virtuously, and one will always follow the directions of one’s conscience. Moreover, it is this reason that should motivate one when one acts virtuously: as Butler puts it, virtue consists “in affection to and pursuit of what is right and good, as such” (XI.20). Nonetheless, as a matter of fact, a virtuous course of life will never be certain to make one less happy than any vicious alternative way of life.

Even if this is the correct interpretation of the content of Butler’s claims about the harmony of virtue and self-interest, we still have to deal with the second of the two exegetical problems that I identified above: How exactly does Butler think he can argue for this sort of harmony between virtue and self-interest?

7. Butler’s arguments for his claims about virtue and self-interest

Butler presents his arguments for this sort of harmony between virtue and self-interest several times — for example, in Sermon I (I.14), and in Sermon III (III.7–8). These arguments are based on his fundamental conception of self-interest, which is first briefly explained in the Preface (P 35–42) and in Sermon I (I.7), but receives its fullest exposition in Sermon XI.
Butler is particularly insistent that self-love, the steady calculating desire that each person has for his own self-interest or happiness, must be distinguished from the particular passions. The distinction is illustrated by a striking pair of examples (I.7, note):

And as self-love and the several particular passions and appetites are in themselves totally different; so, that some actions proceed from one, and some from the other, will be manifest to any who will observe the two following very supposable cases. One man rushes upon certain ruin for the gratification of a present desire: nobody will call the principle of this action self-love. Suppose another man to go through some laborious work upon promise of a great reward, without any distinct knowledge what the reward will be: this course of action cannot be ascribed to any particular passion. The former of these actions is plainly to be imputed to some particular passion or affection, the latter as plainly to the general affection or principle of self-love.

He analyses this difference in the following way (XI.5):

Every man hath a general desire of his own happiness; and likewise a variety of particular affections, passions, and appetites to particular external objects. … The object the former pursues is somewhat internal, our own happiness, enjoyment, satisfaction; whether we have, or have not, a distinct particular perception what it is, or wherein it consists: the objects of the latter are this or that particular external thing, which the affections tend towards, and of which it hath always a particular idea or perception. The principle we call self-love never seeks any thing external for the sake of the thing, but only as a means of happiness or good: particular affections rest in the external things themselves. One belongs to man as a reasonable creature reflecting upon his own interest or happiness. The other, though quite distinct from reason, are as much a part of human nature.

Here Butler insists that the object of self-love is one’s own happiness as such; self-love need not involve any particular “perception” of what concrete external goods one’s happiness will involve, as is shown by the case of the man who toils laboriously “upon promise of a great reward” without knowing what the reward will be. Moreover, he insists that one’s happiness is something “internal”: that is, it is purely a mental state, not a state of the external world. Self-love seeks particular external goods only as a means to this internal state of happiness.

By contrast, the objects of the particular passions are “external objects”, which are
desired simply for their own sake, and not merely as means to happiness or pleasure. He offers a famous but obscure argument for the conclusion that the particular passions must be conceived in this way (XI.6):

That all particular appetites and passions are towards external things themselves, distinct from the pleasure arising from them, is manifested from hence; that there could not be this pleasure, were it not for that prior suitableness between the object and the passion: there could be no enjoyment or delight from one thing more than another, from eating food more than from swallowing a stone, if there were not an affection or appetite to one thing more than another.

Butler is sometimes interpreted here as identifying pleasure or enjoyment or delight with the satisfaction of a desire. Now, it certainly seems true that if there are to be any satisfied desires at all, there must be desires for something other than just for the satisfaction of desires as such. So if this identification of pleasure with the satisfaction of desire is correct, then Butler’s argument is sound.

However, this identification of pleasure with the satisfaction of desire seems clearly incorrect. First, one can desire things of which one will never have any knowledge: for example, one might desire to be remembered after one’s death; but even though this desire is satisfied (one is remembered after one’s death), the satisfaction of this desire does not give one any pleasure since after death, one is no longer capable of pleasure at all. Secondly, even if one does know of the satisfaction of one’s desire, this knowledge might just fail to bring any pleasure at all. Finally, pleasures can take one by surprise: one may just suddenly find oneself taking pleasure in something, even though one never desired it before one starts to feel the pleasure.

Fortunately, we do not have to interpret Butler in this way. In objecting to the

identification of pleasure with the satisfaction of desire, I assumed that “desire” is a mental state that we have towards objects that we do not know to have been attained (such as objects that are absent or uncertain, or can be attained only in the future).\(^{32}\) But Butler’s use of the terms ‘affection’ and ‘passion’ need not be restricted in this way. It is plausible that these terms refer simply to any mental state or mental disposition that involves an element of feeling. When Butler speaks of an “affection for” a certain object, he means a state that involves a positive feeling that is directed towards that object. So when he says “the very idea of interest or happiness consists in this, that an appetite or affection enjoys its object” (P 37), he is saying that pleasure or happiness consists in any sort of positive feeling that is directed towards an object that one knows to have been attained.

Butler’s basic insight, I believe, is that pleasure has an (“intentional”) object, and that one’s pleasure is in a sense “based on” the knowledge that this object has been attained. When we are pleased that \(p\) is the case, our pleasure is based on the knowledge that \(p\) is the case; when we are pleased by the sound of beautiful music, we know that we are hearing the music; and so on. Now, sometimes one might be pleased merely at the thought that one will be pleased by something or other. But this cannot always be the case: sometimes, the object of one’s pleasure must be some object other than one’s own pleasure. In short, it must sometimes be the case that the pleasure is a passion or affection towards a particular external object (specifically, a particular external object that one knows to have been attained).

Many of our pleasures, then, we get from the satisfaction of affections that are directed towards objects other than our own pleasure itself. Now, when all of the elements of our nature are functioning properly (within “their natural stint and bound”), we will still have

\(^{32}\) As Plato puts it, less precisely though more intuitively, we desire things that we lack, not things that we have (Symposium 200a).
a great many particular passions and affections. To focus on the particular passions and affections that are discussed most extensively in the later sermons: (i) we will feel a moderate delight in talking and chatting with other people; (ii) we will feel compassion towards those who are suffering; (iii) we will feel moderate resentment towards injustice; and above all (iv) we will feel benevolent good-will towards other people. But then whenever we know that the objects of these affections have been attained we will feel a commensurate pleasure or satisfaction. Thus, we will feel pleased when we know (i) that we are chatting communicatively about topics of mutual interest with other people, or (ii) that suffering has been relieved, or (iii) that those have violated the rights of others have been brought to justice, or (iv) that other people’s interest or happiness has been promoted.

Moreover, when all the elements of human nature are functioning properly, in addition to these particular passions, we will also have an affection towards virtue as such: we will want to be virtuous, and will take pleasure in the exercise of virtue, purely for its own sake. Thus, when we have the attitudes and dispositions that are characteristic of virtuous people, virtuous conduct will also bring pleasure. This emerges most clearly in the passage where Butler distinguishes between benevolence “considered as a natural affection” and benevolence “considered as a virtuous principle” (XI.16):

Happiness consists in the gratification of certain affections, appetites, passions, with objects, which are by nature adapted to them. … Love of our neighbour is one of those affections. This, considered as a virtuous principle, is gratified by a consciousness of endeavouuring to promote the good of others; but considered as a natural affection, its gratification consists in the actual accomplishment of this endeavour. Now indulgence or gratification of this affection, whether in that consciousness or this accomplishment, has the same respect to interest, as indulgence of any other affection …. Thus it appears, that benevolence and the pursuit of public good hath at least as great respect to self-love and the pursuit of private good, as any other particular passions, and their respective pursuits.

The “natural affection” of benevolence is simply the wish to promote the good of others; the
“virtuous principle” of benevolence involves a tendency to be pleased by the consciousness of one’s own endeavours to promote the good of others.

Since there is such a range of pleasures that are characteristic of the virtuous way of life, Butler thinks that there is no reason at all to think that these pleasures will be any less than those of the various vicious ways of life. He tries to make this point plausible by giving some rather anecdotal reflections on the unhappiness that tends to go along with certain forms of vice. For example, he enumerates the disadvantages of lives that are characterized by excessive covetousness, ambition or intemperance (I.14):

that persons in the greatest affluence of fortune are no happier than such as have only a competency; that the cares and disappointments of ambition for the most part far exceed the satisfactions of it; as also the miserable intervals of intemperance and excess, and the many untimely deaths occasioned by a dissolute course of life: these things are all seen, acknowledged, by every one acknowledged.

He also makes similar observations about the drawbacks of lives that are characterized by envy, rage and (excessive) resentment (III.8):

Let it not be taken for granted that the temper of envy, rage, resentment, yields greater delight than meekness, forgiveness, compassion, and good-will; especially when it is acknowledged that rage, envy, resentment, are in themselves mere misery; and the satisfaction arising from the indulgence of them is little more than relief from that misery; whereas the temper of compassion and benevolence is itself delightful; and the indulgence of it, by doing good, affords new positive delight and enjoyment.

Thus, if we compare these different “tempers”, there is no reason to think that the expected benefits of the temper of virtue are any less than those of the various tempers of vice.

Moreover, once we have the temper that is characteristic of the virtuous, then performing a particular vicious action will secure us pleasures that will seem paltry and insignificant compared to the inevitable pains of self-condemnation and self-dislike (P 28):
one may appeal even to interest and self-love, and ask, since from man’s nature, condition, and the shortness of life, so little, so very little indeed, can possibly in any case be gained by vice; whether it be so prodigious a thing to sacrifice that little to the most intimate of all obligations; and which a man cannot transgress without being self-condemned, and, unless he has corrupted his nature, without real self-dislike: this question, I say, may be asked, even upon supposition that the prospect of a future life were ever so uncertain.

Thus, Butler concludes (III.8–9):

Self-love, though confined to the interest of the present world, does in general perfectly coincide with virtue, and leads to one and the same course of life. But whatever exceptions there are to this, which are much fewer than they are commonly thought, all shall be set right at the final distribution of things. … Duty and interest are perfectly coincident, for the most part in this world, but entirely and in every instance if we take in the future and the whole, this being implied in the notion of a good and perfect administration of things.

In this conclusion, Butler reminds us, parenthetically, of something that he believes but does not rest his argument on — namely, the traditional Christian doctrine of the after-life, which will ensure that in the end, virtue and happiness will coincide, without any of the “exceptions” that occur in this life, “entirely and in every instance”.

8. The evaluation of Butler’s arguments about virtue and happiness

It seems to me that there are pressures arising from morality itself to accept something like Butler’s claims about the general harmony between virtue and happiness. Virtuous parents presumably want to inculcate virtuous dispositions in their children, but they presumably also earnestly hope that their children will have lives characterized by a rich array of pleasures and enjoyments. It would be hard for such parents to accept that they can inculcate virtuous dispositions in their children only at the cost of depriving them of many of the pleasures of
life. R. M. Adams puts this point well:33

[It] is hard to deny the moral importance of believing that the moral life will be good, or is apt to be good, for other people. For it is part of moral virtue to care both about the other person’s good and about the other person’s virtue. Morality requires that we encourage each other to live morally. But how could we do that in good conscience if we thought living morally would be bad for the other person? … So it seems that, if we do not believe that living morally is at least normally good for a person, there will be a conflict in the very soul of morality that threatens to tear it apart.

It seems highly plausible to me that there is an important insight behind what Adams is saying here. But it is not clear that Adams’s statement of this insight is exactly right. Perhaps all that morality requires of us is that we should not harm other people, and beyond that basic requirement of non-maleficence, we should not attempt to promote their happiness in any way that might undermine their virtue. Perhaps we are also not required to encourage others to achieve the sort of supererogatory virtue that is characteristic of the saint or the hero. So perhaps all that it is morally important to believe is that people are not normally harmed by complying with what morality strictly requires of them.

Be that as it may, Butler’s arguments for the harmony of virtue and self-interest seem to me pure wishful thinking. The anecdotal evidence that he adduces about the troubles of the various vicious ways of life provides only the flimsiest support to his conclusion. Moreover, a very little reflection will reveal the implausibility of what he says. Is it really true that “the temper of compassion and benevolence is itself delightful”? Butler himself takes the Biblical passage where we are told to “weep with them that weep” (Romans xii.15) as the epigraph to his two sermons on compassion; and he surely cannot have forgotten that the man whom he regarded as his Lord and Saviour “wept for Jerusalem” (Matthew 23.33–39), apparently out of a sense of compassion for suffering and dismay at injustice.

Butler focuses on the extravagant forms of vice (intemperance, covetousness, and excessive ambition) which are no doubt very often conjoined with unhappiness. But he neglects one very common form of vice, which consists simply in a callous attitude towards those who are sufficiently distant from us that it is easy for us not to “feel the effects of their resentment”\(^\text{34}\) — such as the poor and the oppressed in faraway countries of which we know little, or the future generations who will have to deal with the environmental degradation that we have left behind us. It is hard to see how this sort of vice deprives us of many pleasures, but easy to see how it will spare us the anxiety that more virtuous people will feel.

In short, Butler’s claims about the harmony of virtue and self-interest seem implausible to me. Nonetheless, his claims about our “obligation to the practice of virtue” seem to me essentially correct. We must, I think, face the hard fact that a virtuous life is the right or proper life for us to lead — even though by living such a life we expose ourselves to various sources of pain and anxiety that the vice of callousness would spare us from.

\(^{34}\) The phrase is due to David Hume, *Enquiry concerning the Principles of Morals*, Section III (“Of Justice”), Part I.