Obligation, Justice, and the Will in Hume's Moral Philosophy
Margaret Watkins Tate
Hume Studies Volume 31, Number 1, (2005) 93 - 122.

HUME STUDIES
Volume 31, Number 1, April 2005, pp. 93–122

Obligation, Justice, and the Will in Hume's Moral Philosophy

MARGARET WATKINS TATE

I. Introduction

Recent Hume scholarship has shifted from the traditional view that Hume's account of the will and practical reason directly opposes Kantian rationalism about morals. Some scholars now find common ground between Hume's motivational psychology and Kantian understandings of reason and obligation. Although this trend has corrected certain misreadings of Hume as, for example, a straightforward subjectivist, it goes too far in other respects. In the following, I argue that we can understand one aspect of Hume's study of morals—his explanation of the artificial virtue of justice—in a way that avoids such mistakes. I begin by considering Stephen Darwall's argument, in *The British Moralists and the Internal 'Ought,'* that features of Hume's account of justice reveal both an inadequacy in the empirical naturalist tradition of which Hume is the hero and underlying commitments to the proto-Kantian tradition.[1] I suggest that the puzzles about Humean justice identified by Darwall can be dissolved by calling on broader aspects of Hume's ethics and reinterpreting crucial passages in his discussions of justice. I defend Hume against the charge of inconsistency by suggesting an alternative interpretation of Hume's theory of the will and his arguments about the development of justice as a virtue. Finally, I argue that Hume's theory of the will can consistently account for motives to Humean justice, properly understood.

Margaret Watkins Tate is Assistant Professor of Philosophy, Baylor University, PO Box 97273, Waco, TX 76798-7273, USA.
e-mail: Margaret_Tate@baylor.edu

In the *Treatise*, Hume says: "'Tis from the prospect of pain or pleasure that the aversion or propensity arises towards any object: And these emotions extend themselves to the causes and effects of that object, as they are pointed out to us by reason and experience" (T 2.3.3.3; SBN 414).[2] When we act on this propensity or aversion, we say that we exercise our will, defined as "nothing but *the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body, or new perception of our mind*" (T 2.3.1.2; SBN 399). "The will exerts itself," he tells us later, "when either the good or the absence of the evil may be attain'd by any action of the mind or body" (T 2.3.9.7; SBN 439). In the next line, he suggests that we may interchange for "good" and "evil" the words "pleasure" and "pain."[3]

These statements reflect a theory of the will shared by Hume with other early modern British moralists in what Darwall calls the *empirical naturalist* tradition.[4] Empirical naturalists—Hobbes, Cumberland, Hutcheson, Hume, and Locke—share a "desire to account for normativity in fully natural terms, without reliance on supernatural posits and without attributing to reason any powers beyond those involved in empirical inquiry."[5] According to this view, reason may be theoretical and instrumental, but not practical; it neither gives us ends nor motivates.[6] On the other hand, *autonomous internalists*—Cudworth, Shaftesbury, Butler, and (sometimes) Locke—maintain that "*obligation consists in conclusive motives raised through the exercise of autonomous practical reasoning* (that is, the practical reasoning that realizes autonomy.)"[7] These thinkers anticipate ideas that inspire Kant's explicit identification of the self with self-legislating pure practical reason.

As hero of the empirical naturalist tradition, Hume occupies a crucial position in Darwall's study of the development of these competing understandings of normativity. Darwall portrays him as an insightful thinker—so insightful that he recognizes features of justice that he cannot explain in his home tradition's terms. In the end, Hume must "avail himself of the idea that agents can choose an action not because that action may be instrumental in achieving natural goods, but because it is mandated by a . . . normative principle they accept."[8] In availing himself of this idea, Hume espouses an understanding of normativity closer to the ideas of the opposing tradition than his own. If this interpretation of Hume's account of justice is correct, we must revise our understanding of the history of moral philosophy, by placing important aspects of Hume's work in a tradition that culminates with rather than provides an alternative to Kant's moral philosophy.

These issues have implications of interest to contemporary moral philosophy as well. Disagreements between these two traditions continue, though the lines may no longer be clearly drawn.[9] Current debates over the proper account of normativity turn on disputes over the nature of practical rationality. If Hume recognized aspects of justice that conflict with the tenets of his tradition, then

Darwall has in effect given strong evidence, though not conclusive, for the inadequacy of empirical naturalism. The Kantian must still give a coherent account of the kind of normativity the empirical naturalist allegedly cannot explain. But the burden of proof will have shifted: the Kantian at least attempts to explain what some might think is a fundamental element of our moral experience; the Humean, as Darwall suggests, may just be "explaining normativity away."[10] I will not address the question of which tradition explains normativity better or whether it is a mistake to try to "explain normativity away." But clearly, much rides on whether Hume's attempts to explain justice implicitly draw on aspects of an opposing philosophical programme.

II. Pleasure and the Will

Darwall begins by claiming that Hume's theory of the will is instrumentalist (with regard to reason), naturalistic, hedonistic, and egoistic. The defining aspect of this theory is the view that an agent acts if and only if she believes that the action will help her achieve "natural goods." The problem is that the virtue of justice does not seem explicable in these terms. What natural good does the agent pursue when returning property to an owner she knows to be stingy and careless, simply because it is that person's rightful possession? Hume needs to account for such a motive—characteristic of the virtue of justice—and he suggests that the sense of duty or obligation properly motivates just acts. But he also says that actions are virtuous only because of their motives, which therefore must be defined independently of the honor or obligation attaching to particular acts. So how can the motive to just acts simply be our sense that those acts are just? There must be some other motive, Darwall concludes, and he argues that the only available one for Hume is the agent-state of rule regulation or rule obligation, in which an agent takes a rule to bind her regardless of whether she attains or loses natural good by the action. But this motive is inconsistent with Hume's theory of the will.

Most of Darwall's points about Humean justice do not depend on the theory of the will that he attributes to Hume (which is not surprising, given that he thinks these two aspects of Hume's work conflict). Since the charge of inconsistency rests on this theory, however, we must first consider whether some alternative understanding of Hume's account of the will might be more plausible. I will argue that although Hume's theory of the will *is* naturalistic, it is not unambiguously either hedonistic or egoistic. Moreover, Hume's account of how we come to regard certain goods *as* goods or as things worthy of pursuit is more complicated than might appear at first glance.

His account of the will is instrumentalist because Hume denies that reason alone can give agents ends for action; reason enters into deliberation by informing us which means are likely to achieve desired ends.[11] More interesting are the

attributions of hedonism and egoism. The hedonism comes from Hume's repeated equations of good and evil with pleasure and pain, but the egoism is more difficult to locate. After quoting Hume's claim that the will exerts itself when good (pleasure) or evil (pain) may be attained by the mind or body, Darwall says, "Hume's theory of action thus not only employs the traditional idea that the will invariably aims at the good; it also interprets that idea hedonistically and egoistically. Desires and aversions arise from the prospect of pleasure or pain, respectively, for the agent."[12] But this evidence is insufficient to support the charge that Hume's theory is egoistic. Whether or not an agent is egoistic depends upon what kinds of pleasures and pains she has and the relationship between those pleasures and pains and the ends of her actions. Darwall reads Hume as claiming that the agent pursues pleasure as her own good or the satisfaction of her own desires under those descriptions. In other words, if I desire to have Italian food for dinner, I go out to an Italian restaurant because I want the pleasure that I can get from eating Italian food. That pleasure itself becomes my motive. If I invite my friend to join me because I know that she also enjoys Italian cuisine, the end of that action will be something like "to satisfy my desire to give my friend enjoyment," not "to give my friend enjoyment" full stop. When discussing Hume's example of seeking revenge on an enemy, Darwall says, "If . . . will can exert itself only when a good (pleasure) or the absence of an evil (pain) to the agent is attainable, then my motive can only be something like the satisfaction of my desire to punish, rather than, say, that he deserves it or that this will lay him low (or the judgment that one of these is so)."[13]

Reading Hume in this way becomes immediately problematic, however, as Darwall notes. Directly after his claim that the "WILL exerts itself, when either the good or the absence of the evil may be attain'd by any action of the mind or body," Hume complicates his position:

> Beside good and evil, or in other words, pain and pleasure, the direct passions frequently arise from a natural impulse or instinct, which is perfectly unaccountable. Of this kind is the desire of punishment to our enemies, and of happiness to our friends; hunger, lust, and a few other bodily appetites. These passions, properly speaking, produce good and evil, and proceed not from them, like the other affections. (T 2.3.9.8; SBN 439)

Hume seems to say here that we can desire certain ends—the happiness of our friends, for example—simply as such and not because that happiness gives us pleasure, although having that desire fulfilled will of course give us pleasure.

Things become even more complicated in the appendices to the *Enquiry*, where Hume treats the question of self-love more thoroughly than in the *Treatise*. Here Hume argues persistently against those who claim "that, whatever affection

one may feel, or imagine he feels for others, no passion is, or can be disinterested; that the most generous friendship, however sincere, is a modification of self-love; and that, even unknown to ourselves, we seek only our own gratification, while we appear the most deeply engaged in schemes for the liberty and happiness of mankind" (EPM App. 2.2; SBN 296). Darwall's reading of Hume from the *Treatise* implies that Hume himself claims that "we seek only our own gratification." In the *Enquiry*, Hume repeats his claims from the *Treatise* that there are passions such as hunger and vengeance that have a special relationship to pleasure and pain. Here he says that these passions "carry us directly to seek possession of the object" and impel us "immediately to seek particular objects" (EPM App. 2.12; SBN 301). But Darwall claims that if Hume allows that the will "can aim directly at an enemy's (deserved) unhappiness," then he "will be committed to revising his theory of action and the will quite independently of anything he says about the just person's regulation of her conduct by conventionally established rules of property, transfer, and promise."[14]

Is it appropriate to call observations made directly after the statements that Darwall accepts as Hume's theory of the will a *revision* of that theory? Perhaps the theory was never so simplistic. We must admit that the text is ambiguous on the issue, but the very ambiguity means that what gets classed under Hume's theory of the will is a matter for debate. What does Hume mean in the *Treatise* when he says that passions such as lust and the desire for happiness to our friends "produce good and evil" rather than proceeding from them? He seems to elaborate this point in an argument that may be familiar to readers of Bishop Butler:

> Nature must, by the internal frame and constitution of the mind, give an original propensity to fame, ere we can reap any pleasure from that acquisition, or pursue it from motives of self-love, and a desire of happiness. If I have no vanity, I take no delight in praise: if I be void of ambition, power gives me no enjoyment: if I be not angry, the punishment of an adversary is totally indifferent to me. In all these cases, there is a passion which points immediately to the object, and constitutes it our good or happiness; as there are other secondary passions, which afterwards arise, and pursue it as part of our happiness, when once it is constituted such by our original affections. Were there no appetite of any kind antecedent to self-love, that propensity could scarcely ever exert itself; because we should, in that case, have felt few and slender pains or pleasures, and have little misery or happiness to avoid or pursue. (EPM App. 2.12; SBN 301)

Hume's considered view, then, seems to be that although an agent can pursue pleasure or pain as such, she can also pursue other objects for their own sake;

indeed, that the second sort of pursuit may be necessary for the first to be intelligible. It is still true that "the will exerts itself when either the good or the absence of the evil may be attain'd by any action of the mind or body," because the natural relationship between the original passions and their objects makes those objects good or evil. Some things we find desirable because they are pleasant; others we find pleasant because they are desired.

Given these passages, it is difficult to know how to make sense of Hume's remarks in the *Treatise* that "'Tis from the prospect of pain or pleasure that the aversion or propensity arises towards any object" and that "Nature has implanted in the human mind a perception of good or evil, or in other words, of pain and pleasure, as the chief spring and moving principle of all its actions" (T 2.3.3.3,1.3.10.2; SBN 414,118). Perhaps Hume gets carried away in these passages—caught up in the debate over the relationship between reason and the passions to such a degree that he simplifies all of the passions into the desires for pleasure and the abhorrence of pain. Or perhaps he changes his view, although this interpretation is unsatisfying given the proximity of the remarks in "Of the direct passions" that reflect the earlier passages to those that introduce the more complex view. Perhaps, on the other hand, Hume means that pleasure really is the "spring" that actuates movement of the will. When we act, that some object causes pleasure allows us to denominate it "good," and we then pursue that object as a good. So, if I am seeking the pleasure of a summer day, and I know that sitting on my porch swing will result in such pleasure, then I desire to sit on my porch swing and believe that it is good to do so. But if I desire revenge, that passion makes me denominate punishment of my enemy as pleasurable and good (despite the inherent unpleasantness of watching a fellow human suffer), which will then lead to a desire to punish my enemy. Something like this interpretation is suggested by the introduction to the *Dissertation on the Passions*, published after both the *Treatise* and the *Enquiry*:

> Some objects produce immediately an agreeable sensation, by the original structure of our organs, and are thence denominated GOOD; as others, from their immediate disagreeable sensation, acquire the appellation of EVIL. Thus moderate warmth is agreeable and good; excessive heat painful and evil.
>
> Some objects again, by being naturally conformable or contrary to passion, excite an agreeable or painful sensation; and are thence called Good or Evil. The punishment of an adversary, by gratifying revenge, is good; the sickness of a companion, by affecting friendship, is evil.[15]

This interpretation allows that pleasure or pain is always present in intentional action but does not require supposing that Hume always interprets such action

egoistically. We could certainly wish that Hume had discussed the relationship between pleasure and motivation more clearly. It is clear, however, that he did not intend to imply that agents always and only pursue the satisfaction of their desires as such or their own pleasures and pains. Even in the *Treatise*, he is willing to say that natural impulses "frequently" give rise to direct passions for objects that are not independently pleasant or agreeable. It is also clear that Hume accepts what Darwall calls the "traditional idea that the will invariably aims at the good;" the debate is over how things come to be considered good. Darwall's main contention is that Hume's account of justice contradicts this naturalistic aspect of Hume's theory—that the will invariably aims at goods, although the alleged egoism and hedonism present added challenges. So, although there is insufficient evidence to call egoism or hedonism Hume's "official theory of the will," Hume needs a naturalistic account of justice to escape the charge of internal inconsistency.

III. The Motive to Justice

Darwall's argument proceeds in two stages. First, he attempts to establish that Hume employs a sense of obligation—rule obligation—which is distinct from the categories of obligation he explicitly names. Second, he argues that rule obligation must be the motive uniquely approved as the Humean virtue of justice, and that this motive requires practical reasoning not aimed at satisfaction of a desire for some natural good. This exercise of practical reason, he alleges, is more consistent with autonomist internalism than empirical naturalism.

Hume mentions explicitly two kinds of obligation—natural obligation and moral obligation. As J. B. Schneewind observes, Hume follows Shaftesbury in using "obligation" to mean a "determining motive."[16] Darwall takes "natural obligation" to be equivalent to "self-interest." Moral obligation, on the other hand, arises from the approval or disapproval of spectators, which Hume identifies as the source of our distinctions between virtue and vice. Through various mechanisms—including sympathy and the double relation of impressions and ideas—these sentiments obligate agents to perform acts that a virtuous person would perform.[17] So Hume calls the moral obligation to justice the "sentiment of right and wrong:" "The *natural* obligation to justice, *viz*. interest, has been fully explain'd; but as to the *moral* obligation, or the sentiment of right and wrong, 'twill first be requisite to examine the natural virtues, before we can give a full and satisfactory account of it" (T 3.2.2.23; SBN 498). For the sake of brevity, I will not discuss Darwall's arguments for introducing his third concept of obligation.[18] He concludes, however, that Hume needs such a concept, and it must be "a notion of obligation defined through the rules of justice (or their acceptance) themselves."[19] Darwall calls this third kind of obligation "rule obligation" and argues that rule obligation rather than moral obligation is the motive of the

Humean just agent. Unfortunately for Hume, it is also a motive inconsistent with a naturalistic theory of the will.

Hume's difficulties in explaining justice as a virtue arise from the interaction between three elements of that explanation: (1) his claim that acts derive their moral status from motives (2) his argument that the "first virtuous motive" of an act can never be merely the virtue of the act itself and (3) his argument that the merit of just acts is the only apparent motive to justice. (1) and (2) are significantly distinct propositions, although Hume argues that (2) follows from (1). (1) says that we judge an act virtuous or vicious because of its motive; (2) states that the "first motive" behind a virtuous act cannot simply be that the act is virtuous.[20]

Hume commits himself to the principle that we evaluate acts on the basis of their motives prior to his discussion of justice. He lays the groundwork for this claim in Book II in his discussions of love and hatred, and liberty and necessity (T 2.2.3.4,2.3.2.6; SBN 348,410). When he begins his examination of justice, then, he simply states, "'Tis evident, that when we praise any actions, we regard only the motives that produc'd them, and consider the actions as signs or indications of certain principles in the mind and temper" and reflects that we withdraw blame in cases where circumstances beyond a virtuous agent's control preclude the performance of a required act (T 3.2.1.2–3; SBN 477).[21] Having reminded the reader of this fundamental principle, he goes on to give the argument known as "Hume's circle." Since we evaluate primarily character traits or motives, not actions, regard for the virtue of an action requires that its motive has already been deemed virtuous. In Hume's words—

> To suppose, that the mere regard to the virtue of the action, may be the first motive, which produc'd the action, and render'd it virtuous, is to reason in a circle. Before we can have such a regard, the action must be really virtuous; and this virtue must be deriv'd from some virtuous motive: And consequently the virtuous motive must be different from the regard to the virtue of the action. A virtuous motive is requisite to render an action virtuous. An action must be virtuous, before we can have a regard to its virtue. Some virtuous motive, therefore, must be antecedent to that regard. (T 3.2.1.4; SBN 478)

The circle argument seems to say that no action can be virtuous unless there is some motive other than the virtue of the action itself or the sense that the action is morally obligatory. The response to the question about some act, x, "Why is x virtuous?" is always, "Because x is done from a virtuous motive." The response to the question, "What is that motive?" cannot be "Because x is virtuous." Such a response ensnares us in vicious circularity.

It appears, then, that if justice is to be a Humean virtue, the just person must act from some motive of which a spectator could approve other than the conviction that the act she performs is just and just acts are obligatory. Indeed, we have two questions: What good is the just agent pursuing, so that her motive is consistent with Hume's theory of the will? And what character trait could be associated with that motive, so that a person lacking that trait would be condemned by observers in ways that would generate a sense of obligation to be just? On Hume's view, virtue requires a relatively fixed disposition of temper, since "[a]ctions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil" (T 2.3.2.6; SBN 410).

In the course of arguing that justice is an artificial virtue, Hume examines possible motives to just acts, distinct from regard to justice itself, that would engender just actions when required. Although he dismisses self-love as "the source of all injustice and violence" and rejects it along with public interest and private benevolence as motives to just acts in a state of nature, he argues that corrected self-interest naturally obliges the establishment of the rules of justice (T 3.2.1.10; SBN 480). Justice is required to avoid a more tame Hobbesian state of nature. So, Darwall considers whether self-interest might be the natural good sought by the just agent.

The problems with this suggestion are familiar: justice is invariant, but consequences are not. If interest is the motive to justice, it must provide such a motive even in cases where an agent might benefit more by transgressing the rules. At least in the *Treatise*, Hume provides evidence that he accepts that each individual act of justice either produces good or upholds the practice as a whole, which is necessary to each agent's good. Although he may admit the possibility of free-rider cases as long as one ignores the consequences of injustice for the practice of justice, he insists that those potential consequences create a situation in which "every individual person must find himself a gainer, on ballancing the account" (T 3.2.2.22; SBN 497). As Darwall says, "The 'singular connection' [between justice and interest] is that it is only because others are likely to comply only if one does—that their compliance with rules of justice is among the consequences of one's compliance—that each individual agent's complying with the rules of justice invariably has the best consequences."[22]

Suppose Hume's view in the *Treatise* is that interest always provides a motive to just acts, provided that we take into account the fragility of the convention. Following a rule for the sake of a "natural good" is not equivalent to following it because it is a rule. The latter gives the rules "independent deliberative weight," which is, Darwall claims, inconsistent with Hume's theory of the will.[23] But on this interpretation, the just agent acts for her own interest rather than

considering the rules intrinsically motivating. Thus, Darwall concludes that the *Treatise* allows for the possibility of a theory of justice that does not conflict with empirical naturalism.

One serious problem with this interpretation, however, is that Hume continually insists, particularly in his discussion of the artificial virtues, that human nature does not work this way. Just as we fixate on our own goods and those of our friends and family, sometimes to the expense of other duties, we fixate on present and closely available goods, often at the expense of what might be best for us in the long run. He notes that, "The consequences of every breach of equity seem to lie very remote, and are not able to counter-ballance any immediate advantage, that may be reap'd from it" (T 3.2.7.3; SBN 535).[24] The concrete circumstances of life include a scarcity of resources presenting endless temptation, an ever-increasing society contributing to a sense of anonymity and hence irresponsibility, and constant seduction by "more present" interests (cf. T 3.2.2.24; SBN 499). Although "on the first formation of society," people are moved "[t]o the imposition . . . and observance of these rules, both in general, and in every particular instance" by "a regard to interest," interest will not do as a motive to justice in a developed nation.

The system of justice works only if I, as a member of society, can expect that enough people will abide by its rules that I need not be in constant fear of being a "cully of my integrity, if I alone shou'd impose on myself a severe restraint amidst the licentiousness of others" (T 3.2.7.3; SBN 535). Darwall notes that increased population complicates matters for Hume: "When society becomes sufficiently numerous, individuals may tend to lose sight of their real interests, but these continue to dictate abiding by the rules of justice in every case."[25] But as more members of society lose sight of their interests, the less it is in anyone's interest to abide by the rules of justice. If enough people flout the rules, the "singular connection" between justice and interest is severed. The rare just citizen, if her only motive is self-interest, is then a fool. The system of justice works only if it is stable, which it will not be if self-interest is the only motive to justice. Each act of injustice, Hume says, "affords me a new reason for any breach of equity" (T 3.2.7.3; SBN 535).[26] If large numbers of people flout the rules, justice-as-self-interest would prescribe transgressing the rules. But it is essential to justice that it is "universal and perfectly inflexible" (T 3.2.6.9; SBN 532). Thus, in a developed nation, it would fail too often, and then its foundations would be undermined.

This argument suggests that, even with the "singular connection" between interest and the public expression of that interest as manifested in the conventions of justice, self-interest cannot serve as the motive for justice. Moreover, we have reason to believe Hume recognizes this problem. As Darwall notes, Hume suggests that the "sense of duty" not only provides necessary additional motivation but is the appropriate motive for abiding by the rules of justice. For example, consider Hume's description of the intercourse of civilized men:

> I suppose a person to have lent me a sum of money, on condition that it be restor'd in a few days; and also suppose, that after the expiration of the term agreed on, he demands the sum: I ask, *What reason or motive have I to restore the money?* It will, perhaps, be said, that my regard to justice, and abhorrence of villainy and knavery, are sufficient reasons for me, if I have the least grain of honesty, or sense of duty and obligation. (T 3.2.1.9; SBN 479)

That one's sense of duty and obligation is the appropriate motive for "man in his civiliz'd state . . . when train'd up according to a certain discipline and education" is one of the distinguishing characteristics of the artificial virtues (T 3.2.1.9; SBN 479).[27]

Recall, however, that the circle argument seems to establish that the sense of duty can never be the only motive to virtuous acts, unless some other admirable motive can be associated with the act. How, then, could an obligation to just acts be generated? These considerations bring us to our second question: what character trait could be associated with justice, so that a person lacking that trait would be condemned by observers in ways that would generate a sense of obligation to be just? The agent cannot be moved solely by the belief that the actions in question are just, if their being just depends on their causal relationship with some motive antecedently defined as just. Moreover, the motives must be associated with some stable motivational states, since moral disapprobation generates motives because it relates to enduring character traits and pronounces judgments on persons. Strength of mind—the disposition to resist the temptation of present and near interests for one's own greater and more distant interest—might serve the purpose in the case of justice, Darwall argues.[28] After all, self-interest is the natural obligation to justice, and it must be corrected by virtue to make up for our tendency to sacrifice distant interest to short-term gains. A society might subsist with a system that encouraged people to counteract this natural tendency through approval of the disposition to follow one's calm passions rather than the violent.[29] Darwall concludes that, in the *Treatise*, Hume's claims are consistent with the view that we approve just persons because of their strength of mind: "their motive is always given by enlightened self-interest."[30] Darwall grants that this interpretation stretches the text in several ways. One of the more serious difficulties with this proposal is that it makes it difficult to understand why we would distinguish justice from other traits that employ strength of mind, such as prudence.[31] And it certainly does not cohere with Hume's suggestions that moral obligation is the proper motive to justice in a civilized state. Nevertheless, Darwall believes that the considerations presented by Hume's circle argument make suitably corrected interest the most likely candidate for the motive approved of as justice.

Of course, enlightened self-interest cannot serve this function if there are cases of genuine free-rider dilemmas—if individual acts of injustice might in some cases (however rare) be in the long-term best interest of an individual agent. And in the "sensible knave" passage from the *Enquiry*, Hume seems to admit that there may be such cases:

> Treating vice with the greatest candour, and making it all possible concessions, we must acknowledge that there is not, in any instance, the smallest pretext for giving it the preference above virtue, with a view to self-interest; except, perhaps, in the case of justice, where a man, taking things in a certain light, may often seem to be a loser by his integrity. And though it is allowed, that, without a regard to property, no society could subsist; yet according to the imperfect way in which human affairs are conducted, a sensible knave, in particular incidents, may think, that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy. That *honesty is the best policy*, may be a good general rule; but it is liable to many exceptions; And he, it may, perhaps, be thought, conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions.
>
> I must confess, that, if a man think, that this reasoning much requires an answer, it will be a little difficult to find any, which will to him appear satisfactory and convincing. If his heart rebel not against such pernicious maxims, if he feel no reluctance to the thoughts of villainy or baseness, he has indeed lost a considerable motive to virtue; and we may expect, that his practice will be answerable to his speculation. (EPM 9.22–23; SBN 282–3)

If the knave is correct, and villainy best serves his interest, then strength of mind cannot be the virtue approved in the just agent. Interested strength of mind would dictate that the knave break the rules in certain cases, but justice, like all artificial virtues, admits of no exceptions.

It is odd, perhaps, to read the sensible knave passage as Hume addressing the free rider problem. He hedges continually throughout the passage, not quite willing to admit that the knave sees his options clearly. Nonetheless, Hume's criticism of the knave depends upon there already being a vice characterized as injustice: the knave will lose "[i]nward peace of mind, consciousness of integrity, a satisfactory review of [his] own conduct," and "the invaluable enjoyment of a character" (EPM 9.23,25; SBN 283). If these moral advantages are necessary for justice to be in our long-term interest, then strength of mind cannot be the

trait originally approved of that generates the distinction between justice and injustice as virtue and vice. We would first have to approve of justice before those advantages would obtain, so acting justly would not be in accord with the agent's interest prior to the invention of the virtue, however adept the agent is at keeping distant effects in mind.[32] These considerations, combined with the other infelicities of conceiving of the motives of prudence and justice as one and the same, eliminate strength of mind as a possibility for the just agent's particular virtue.

Darwall concludes that the only remaining possibility for the agent state unique to justice is a state of mind inconsistent with Hume's theory of the will. The just agent takes the rules as *themselves* normatively authoritative, distinct from any concern for natural good. She accepts the rule as a *norm*, and it moves her as such. She does not regard the rules merely externally, as tools that serve her interested purposes. She accepts them internally, as necessitating actions for her. It is not that she follows the rules because they support a greater good than actions that contravene them. The rules prohibit her performing such an action *regardless* of the weight of goods to be obtained. As Darwall has it, the rules themselves are obligating. "So the relevant virtue—justice as an agent state—is realized when agents act because they so regard the rules—that is, from a sense of (rule) obligation."[33] Because rule obligation thus conceived does not involve aiming at goods, it is inconsistent with Hume's theory of the will.

IV. Revisiting the Circle

Hume can respond to this charge of internal inconsistency, but to construct the response, we must first reconsider Hume's difficult circle argument that plays such an important role in Darwall's interpretation. Part of what makes the circle so puzzling is Hume's own reticence about the connection between this argument and the main topic of the section that it introduces. The argument itself seems to be a general remark about moral motivation, but Hume clearly intends it to support his claim that justice is an *artificial* virtue. But he is not explicit about how the argument supports that claim.

The section in question is entitled "Justice, whether a natural or artificial virtue?" After reminding us that he has already suggested that some virtues might be artificial in the relevant sense, Hume says, "Of this kind I assert *justice* to be; and shall endeavour to defend this opinion by a short, and, I hope, convincing argument, before I examine the nature of the artifice, from which the sense of that virtue is deriv'd" (T 3.2.1.1; SBN 477). This argument consists of a discussion of the principle that "*no action can be virtuous, or morally good, unless there be in human nature some motive to produce it, distinct from the sense of its morality*" (T 3.2.1.7; SBN 479); an examination of various possibilities for such a motive in the case of

justice—self-interest, a regard to public interest, and private benevolence—each of which is rejected in turn; and the explicit conclusion—

> From all this it follows, that we have naturally no real or universal motive for observing the laws of equity, but the very equity and merit of that observance; and as no action can be equitable or meritorious, where it cannot arise from some separate motive, there is here an evident sophistry and reasoning in a circle. Unless, therefore, we will allow, that nature has establish'd a sophistry, and render'd it necessary and unavoidable, we must allow, that the sense of justice and injustice is not deriv'd from nature, but arises artificially, tho' necessarily from education, and human conventions (T 3.2.1.17; SBN 483).

Darwall spends little time discussing the structure of this argument, but he quotes this passage often. Whenever he does, he treats the circle as if it were a dangling problem for Hume—a sign of the confusion that rule obligation and the agent state of rule regulation are supposed to clarify. "Taken literally," he says on page 290, "this says that justice is not simply puzzling but paradoxical." And later on page 315, "Hume confusedly asserts that it is the moral sentiment that uniquely motivates artificially virtuous acts. And that is what lands him in the circle."

Hume does not seem so troubled. Apparently, he believes that the suggestion that justice is an *artificial* virtue solves the problem of the circle. One might expect some lamentations of the paradox he has woven, perhaps echoing the end of the first book of the *Treatise*. Instead, he blithely changes the subject, moving immediately to a "corollary" on our tendency to judge motives with reference to their general preponderance among our fellows. He does not seem to think he has a problem for Darwall to solve. His nonchalance would be unimportant had Darwall claimed to have discovered a subtly hidden paradox deep within the text implied by Hume's difficult ideas. But his paradox is given and highlighted by Hume: "there is here an evident sophistry and reasoning in a circle." Hume's nonchalance, then, is a real exegetical puzzle.

Hume's silence, I believe, tells us that his story of the evolution of justice, whereby natural observations about the benefits of society lead to the development of conventions, which then take on moral significance, is crucial to understanding the motives to justice themselves. The circle does not merely remind us that we evaluate primarily motives rather than acts; it tells us something about what those motives have to be. To paraphrase Hume, perhaps it will appear afterwards that the motives to justice must evolve along with the virtue itself.[34]

Recall that the logic of the circle argument appears simple: the response to the question about some act, x, "Why is x virtuous?" is always, "Because x is done from a virtuous motive." The response to the question, "What is that motive?"

cannot then be "Because x is virtuous." We can move this series of questions back a step to capture better Hume's concern with "the first motive, which *produc'd the action*" (emphasis added) as well as rendering the act virtuous. The response to the question, "Why do x?" may be "Because x is virtuous." Hume acknowledges the possibility of such a motive explicitly at T 3.2.1.8 (SBN 479), but he implies that the response only makes sense in a context where we generally acknowledge the virtue of x-type acts. "When any virtuous motive or principle is common in human nature, a person, who feels devoid of that principle, may hate himself upon that account, and may perform the action without the motive, from a certain sense of duty, in order to acquire by practice, that virtuous principle, or at least, to disguise to himself, as much as possible, his want of it."

"Because x is virtuous" is only an adequate response to "Why do x?" if the agent to whom it is given feels the complex pressures of the relevant moral standards. Consider Hume's example of a father who neglects his child. If he were to ask, "Why should I attend to my child when he is whining and disturbing my afternoon nap?" the response that virtuous people act that way should call to mind this man's failure to live up to common standards, to be the person we expect him to be, not his failure to have his acts fit with eternal correspondences between judgments and acts, for example. If the father is not so moved and demands further explanation, asking, "Why *do* we think virtuous people act that way?" Hume will respond in terms of the virtuous motives—in this case, natural affection—that generally produce such a response, and we are back to our original series of questions and answers that began with "Why is x virtuous?"

Notice however, that the response, "Because x is done from a virtuous motive," fails to give a complete Humean response to this question (Why is x virtuous?) in *two* ways. First, as already noted, it needs to be filled in with a description of that motive that goes beyond an obligation to that act. The identification of this motive, however, does not yet exhaust the explanatory powers of Hume's system. No Humean explanation of why a particular action is called virtuous would be complete without an explanation of the mechanisms of sympathy, according to which spectators sense that a particular trait or motive tends to prove useful or agreeable to its possessor or those with whom that person lives or has commerce. Indeed, this judgment, in some sense, is primary: it is with approval or disapproval that morals enter the world. To have a virtue is *essentially* to have a trait that is approved in a specific way. As Hume says, "these two particulars are to be consider'd as equivalent, with regard to our mental qualities, *virtue* and the power of producing love or pride, *vice* and the power of producing humility or hatred" (T 3.3.1.3; SBN 575).

On more than one occasion, Hume insists that the verdict of the spectator is the last word in distinguishing virtue from vice. Although he seeks the "principles" of moral judgment, and finds in all such judgments mechanisms of sympathy and

the pleasurable or painful effects of motives and traits, he does not claim that the moral judges themselves consciously take account of those pleasures and pains in pronouncing judgment:

> An action, or sentiment, or character is virtuous or vicious; why? because its view causes a pleasure or uneasiness of a particular kind. In giving a reason, therefore, for the pleasure or uneasiness, we sufficiently explain the vice or virtue. To have the sense of virtue, is nothing but to *feel* a satisfaction of a particular kind from the contemplation of a character. The very *feeling* constitutes our praise or admiration. We go no farther; nor do we enquire into the cause of the satisfaction. We do not infer a character to be virtuous, because it pleases: But in feeling that it pleases after such a particular manner, we in effect feel that it is virtuous. (T 3.1.2.3; SBN 471)

And in his discussion of allegiance, he goes so far as to suggest that we cannot be mistaken in our general judgment of virtue and vice:

> The distinction of moral good and evil is founded on the pleasure or pain, which results from the view of any sentiment, or character; and as that pleasure or pain cannot be unknown to the person who feels it, it follows, that there is just so much vice or virtue in any character, as every one places in it, and that 'tis impossible in this particular we can ever be mistaken. (T 3.2.8.8; SBN 547)

One implication of these remarks is that the origin of the distinctions between virtues and vices lies squarely in the eyes of the moral spectator. Virtue is just what the moral spectator approves of, and what the spectator primarily approves of are motives, not actions. She considers actions as signs of virtuous motives and denotes them virtuous as such signs. No matter how we come to regard a particular act as virtuous, once that regard is present, the agent has an additional motive to the performance of acts thus approved; the desire to avoid the pain of humility and a bad reputation becomes a moral obligation.

Let us now reconsider our series of questions. The answer to "Why is x virtuous?" in the concrete circumstances of life, which include a richly developed moral system, might be quite various. "Because x is generous," "because x helps others," and "because x shows a healthy disregard for material possessions" could all be correct answers to the same question. Indeed, they are not mutually exclusive and might each be true of the same act. Consider the first response—that x is generous. Notice that once we agree that generosity is a virtue, this response might appear in our series of questions: Why do x? Because x is virtuous. Why is

x virtuous? Because it is generous. An explanatorily complete response would go on to say that generous acts are those produced by generosity, and generosity is a trait characterized by specific motives that tend to be useful to others (and perhaps agreeable to both the self and others as well). Presumably, however, we have come to agree that certain sorts of actions—those generated by the motives associated with generosity—are virtuous because of this association. And once this happens, we may hate ourselves for failing to do those actions, because we believe that failure signifies a failing in our character.

I believe that, with the circle argument, Hume suggests that our sense of justice develops in a similar but importantly different way. With the natural virtues, the motives or traits themselves define the acts that come to be called virtuous, each of which can be specified in terms of the particular trait that gives rise to the act. With the artificial virtues, however, *conventions* define the acts that come to be denominated as belonging to a particular class of actions—the just ones, for instance—and only then does a certain attitude towards those acts come to considered virtuous. Humankind has a class of "just acts," in other words, before it has a virtue called "justice."

This prior delineation of which acts fit into the category of justice and which are unjust means that when someone asks, "Why do x?" (where x is a just act) the response will be different depending on which stage of civil development has been attained when it is asked. As the conventions of justice are just being established—as conventions, not as moral rules—one available response will be "because x is just," but such a response could only be intended to remind the agent of what prudence requires. Why should I return these borrowed goods to my neighbor? Because we agreed to play by that rule, and I know that if I fail to do so, the system will fail. Next time around, she will not return what I originally had in *my* possession.

Once, however, the sense of virtue has been "annexed" (as Hume says at T. 3.2.2.23; SBN 498) to justice, the response to "Why do x?" where x is a just act may well be, "Because x is just," in the same way that "Because x is generous" can be a reasonable response to such a question. As with specific appeals to the natural virtues, such a response can call to the agent's mind her responsibility to live up to common standards and to be the person we expect her to be. As with generosity, the response that appeals specifically to the virtue in question is not explanatorily complete. I have said that with the natural virtues, a complete response says that the acts in question are those produced by a virtue, which is a trait characterized by specific motives that tend to be useful or agreeable to the self or others. With the artificial virtues, the complete response says that the acts in question are those delineated by a convention, and that following such a convention proves to be useful to the self and/or others. Thus, the way in which we come to distinguish certain acts as required by certain virtues is different for these two kinds of vir-

tues—naturally virtuous actions are distinguished by their corresponding motive, but artificially virtuous actions are distinguished by conventions. Regardless of the type of virtue, however, it remains the case that once we (as spectators) have come to agree that certain sorts of actions are virtuous, we (as agents) may hate ourselves for failing to do those actions, because we believe that failure signifies a failing in our character.

If this interpretation is correct, then it is a mistake to read the circle argument as Hume's own search for a distinct motive that makes certain acts just and therefore virtuous. On the contrary, this brief argument is part of a larger discussion whose primary purpose is to establish that justice is an artificial rather than a natural virtue. The discussion accomplishes this purpose by showing that, in contradistinction to the natural virtues, there is no natural motive to just acts that would be approved by spectators prior to the convention. What is just and unjust must be determined by the conventions themselves, rather than being singled out by the approval or disapproval of natural motives. Once this sorting of acts is accomplished, however, there need be nothing mysterious about the motives of the just person. She believes those acts are morally obligatory, because she believes that they are required by the virtue of justice, and she believes that good people are just.

In other words, Hume has a naturalistic explanation for the motive to just acts both for people who live in nascent societies and for those who live in more developed ones where the rules of justice have been long established. Those in nascent societies are moved by their own self-interest to make rules, and they further see that their interest requires making those rules inflexible. (Allowing justice to follow the usual course of general rules, which take into consideration particulars of circumstance and character, would "produce an infinite confusion in human society" (T 3.2.6.10; SBN 532).) Presumably, the move to seeing such principles as associated with a particular virtue will be a gradual one. As they are passed on, however, the inheritors of the convention will be taught that these actions are required by virtue—specifically, by the virtue of justice. They might have a train of motives leading up to their own acceptance of this view, including desires to avoid sanctions or receive certain rewards.[35] As long as they develop the sense that they are obligated to behave justly, however, they have a moral motive to do so. (Such approval is moral approval, regardless of how it is generated.) It is not a circular motive, because, uniquely with the artificial virtues, the motives themselves do not specify which acts correspond to the virtue; they do not explain why these acts are considered just.

Of course, the mechanisms of moral education cannot on their own tell the entire story of the development of the virtue. Hume needs to explain why, for example, some particular set of actions came to be denominated as stealing (which will of course require a story about the establishment of property regulations) and

then why moral sentiment came to be associated with such acts.[36] As he has told us earlier, what we need to explain is why just acts or the esteem of justice cause their particular pleasure (See T 3.1.2.11; SBN 475–6). At the beginning of the section on justice, he says "that there are some virtues, that produce pleasure and approbation *by means of an artifice or contrivance*, which arises from the circumstances and necessity of mankind" (T 3.2.1.1; SBN 477, emphasis added). Justice is just this kind of virtue, and he must explain why the contrivance arose. His "undoubted maxim, *that no action can be virtuous, or morally good, unless there be in human nature some motive to produce it, distinct from the sense of its morality*" still stands (T 3.2.1.7; SBN 479). Self-interest, suitably corrected, explains why we agree to establish property rights and other rules of justice; interest is thus the natural motive or *natural obligation* to justice. But the genealogical aspect of Hume's account and the claim that justice develops in stages is important here. When the conventions of justice are first established, they have nothing to do with moral motives. There is, in other words, a motive to just acts in human nature distinct from the sense of their morality. Such a sense is only "annexed" to justice once people notice the good effects of such a system and sympathize with the pleasures it produces and the pains involved in transgressions. As society expands, moral motives develop and become the *proper* motives to justice. Even this stage has divisions: at first, the natural mechanisms of sympathy forward the approbation and disapprobation. Later, politicians and parents use moral education to augment these motives. These different stages must be taken into account in any accurate explanation of Hume's theory of justice.[37]

V. Abhorrence of Villainy and Knavery

In this section, I develop my interpretation of Hume's theory of justice by responding to two related objections: first, that my view amounts to the claim that we approve first and foremost of just acts rather than just persons, which is to abandon a central principle of Hume's virtue ethics; and second, that my attempt to respond to this first objection makes my view equivalent to Darwall's, meaning that Hume in fact approaches an autonomous internalist conception of the will.

I have argued that the conventions' delineation of just acts takes the explanatory place that motives play for the natural virtues, that we determine which acts are just by artifice, whereas we determine which acts are, for example, generous, by noticing which acts are commonly produced by generous motives. We then train people to regard those acts as morally obligatory, and doing so engenders in civilized people the belief that good people do not fail to perform such acts. But, the objector insists, here is where the story fails, at least for Hume. To get people to believe that "good people act justly," we must convince them of something other than the mere belief that a set of acts is obligatory. We must associate such acts

with some particular character trait, or they will not see the acts as obligatory at all, at least if Hume is right about the nature of moral obligation.

Insisting, as I have done, that Hume does not intend the circle argument to deliver a unique motive to justice does not solve this problem. The principle that moral approbation is of character traits rather than actions is not established by the circle argument; it is a premise of that argument, and Hume seems committed to it for independent reasons. For example, in his discussion of liberty and necessity, he explains that censure or praise must attach to a person, since, "Actions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon them, and can neither redound to his honour, if good, nor infamy, if evil" (T 2.3.2.6; SBN 411). We need this connection between character and act, he argues, for punishment to make any sense. Unless people associate just acts with character traits and motives, then how can the idea of virtue ever be "annexed" to that of justice? What sense could be made of Hume's claims that regard to one's character requires one to "fix an inviolable law to himself, never, by any temptation, to be induc'd to violate those principles, which are essential to a man of probity and honour" (T 3.2.2.27; SBN 501)? For justice to transform from a merely self-interested convention into a moral institution, it must, according to Hume, be associated with persons and their motives. Hume allows that motives of duty alone may produce action, of course, in the passage at 3.2.1.8 (SBN 479) that explains how someone may "hate himself" for lacking a virtue and perform an action "from a certain sense of duty." Darwall suggests that such instances are "rare," according to Hume,[38] but there is no textual evidence that Hume thought so. But what does the agent hate herself for lacking? What constitutes the virtue in the concrete circumstances of life once that concept has been annexed to justice?

Suppose we say that the just agent feels obligated to be the sort of person who finds just acts obligatory. Then, the objector will press, her sense of obligation is incoherent. To find just acts obligatory is simply to want to be a just person or have the trait of justice, since we only feel a sense of obligation when we believe we are lacking some such trait. But we were looking for an explanation of what constitutes the trait of justice, so it seems we are again moving in a circle.

Perhaps, however, we can fill out the sense of "the sort of person who finds just acts obligatory" in a way that avoids such circularity. Presuming that this is an agent whose psychology Hume would countenance, her finding just acts obligatory cannot consist in her sensing some kind of innate and bare "to-be-doneness" in a proposed course of action. But couldn't the development of the virtue of justice associate certain acts with a certain type of character in such a way that "Because x is just" could be a sufficient reason for acting in a particular way? Consider the following analogy, inspired by Hume's own comparison of the conventions of justice with men working together to row a boat (T 3.2.2.9; SBN 490).

Imagine that Joe is on a crew team and is an excellent rower—the sort of rower held up as an example to his fellow team members and admired by them in turn. He seems to behave in all the ways that an excellent rower does: he senses small changes in the acts of the rest of the crew and responds accordingly (thereby contributing to the likelihood of the team being in "swing"); he responds quickly to the calls of the coxswain; perhaps most importantly, he sees himself as a member of the team and would do almost anything rather than distract the other members by any behavior that might call attention to himself during a competition. Let us further imagine that rowers like Joe have, as the sport of rowing developed, come to be called by a particular term: they are known as "swift."

Rowers who take the sport seriously may see swiftness as an ideal, and they may even become quite angry with rowers who behave in a non-swift manner. After observing the dim effects of such behavior on the other members of the team, they might sympathize with the victims of non-swiftness, and they may even come to feel something approaching moral disgust at those guilty of such behavior. Outside of competition, swift acts have little value whatsoever (although they may be prompted by traits that do have such value; for instance, a lack of arrogance). Inside the competition, they are all-important. Notice that once one's teammates begin to feel the almost-moral disgust at non-swiftness and almost-moral approval of swiftness, Joe has an additional motive to swiftness, besides the self-interested motive of such behavior making it more likely that he will be part of a winning team. He may well feel something approaching moral obligation. He will be chagrined if he fails in his usual swiftness. Of course, there is no such motive apart from the history of the game. But once the rules of competition are in place, and the behaviors associated with swiftness identified, if Joe's little brother, who has just been introduced to the sport, asks "Why should I obey the coxswain?" the answer may well be, "Because it is swift."

Returning to the case of justice, why couldn't Hume conceive of the virtue of justice developing in a similar way? Because of the natural obligation of interest, our ancestors had reasons to refrain from taking things agreed to be one another's possessions, to abide by contracts made, to pay sums of money promised in exchange for certain goods and services. Hume tells us that such behavior follows from rules decided by conventions. After observing how failure to do such things affects others in general, we come to sympathize with the victims of injustice and to feel moral disgust at those who behave in such a manner. Outside of the convention, such feelings would not be possible, because it is only with the convention that just acts become valuable. (Although, as with swiftness, traits that support justice, such as a lack of arrogance, may have more general value.) But once we group such behaviors under a common term, and the tendency to behave this way is approved, one can feel a moral obligation to be a just person, and a sufficient answer to "Why should I return this money?" may well be, "Because it is just."

The point is that we are capable of developing quite complex motives by associating certain behaviors with a virtue term. Perhaps the fully just person is she who not only does what justice the convention requires, but does them because they are what justice requires, and she believes that good people are just. A spectator might well approve of someone who feels that just acts are morally obligatory (who feels that justice is a virtue that good people possess), because such a judgment leads to actions that tend to be useful to the self and others. Since which particular acts qualify as justice is determined not antecedently by the motives of particular persons but by a convention that results originally from self-interested motives, this description is not circular.

But, the objector might press, isn't this person just the one Darwall has described? Even if we can create a notion of a virtue in the way that "swiftness" might be created, what motive does the swift person have other than a respect for certain rules? What good is such a person pursuing? Putting the question in terms of the analogy, however, shows how odd such a proposal is. Of course, early swift rowers are not ultimately motivated by their respect for rules. They instead pursue their interest—an interest whose satisfaction is constituted by being on a well-performing rowing team. We might even bring back the suggestion that strength of mind could be central to the relevant trait, if we understand it to be strength of mind with regard to a particular set of such precisely described interests. Agents who are swift have strength of mind with regard to rowing crew, and just agents (prior to the annexing of virtue to the term) have strength of mind with regard to those interests of theirs connected with property transactions. Once, however, we come to admire the disposition to act justly and consider it part of those virtues that make up the honorable man or woman, these considerations can and should generate a sense of obligation towards those acts that the conventions specify as being just. (Likewise, we can imagine Joe feeling obligated to do swift things and guilty when he fails to do them, because he has failed to live up to the expectations of his role.) At this point, it no longer matters that each individual act of justice in such a developed society may fail to further one's interest, because we have the motive of feeling obligated by our sense that justice belongs to the character of the virtuous. Such an obligation, however, is *moral*. How we describe this motive will depend on our particular understanding of how Humean moral obligation motivates. Perhaps one seeks to avoid the humility of knowing that one's character is vicious, or perhaps those with the virtues consider acts designated as virtuous pleasant or good, and are therefore motivated to pursue them. As long as we can in this gradual way come to associate rule-following behavior with character traits, we have the elements necessary to generate Humean moral obligation. We need not invent the category of rule obligation to account for the motive to justice. The early just agent and the late share the disposition to take justice seriously—the former because she knows

it to be in her own interest, the latter because she knows it to be central to the good person's character.

This explanation seems to cohere with Hume's own description of the motivations of the just. As noted earlier, when he asks what reason one might have for returning money one has borrowed, he says:

> It will, perhaps, be said, that my regard to justice, and abhorrence of villainy and knavery, are sufficient reasons for me, if I have the least grain of honesty, or sense of duty and obligation. And this answer, no doubt, is just and satisfactory to man in his civiliz'd state, and when train'd up according to a certain discipline and education. (T 3.2.1.9; SBN 479)

Later, he notes that responsible parents will inculcate in their children "sentiments of honour" (T 3.2.2.26; SBN 501). Again, in the sensible knave passage, he says that the knave's propensity to "thoughts of villainy and baseness" and his lack of "antipathy to treachery and roguery" indicate that he has lost a "considerable motive to virtue" (EPM 9.23; SBN 283). Hume portrays the just person as one who has genuine respect for others' property, who is appalled by behavior that transgresses the institutions of honor that surround him, and who is likely to identify anyone who fails to share these sentiments as base and vile. In short, he believes that good people are just.

We need not imagine this person as someone unable to account for her respect for justice or her belief that its presence is required in a virtuous character. A Humean agent tempted to injustice has various resources to appeal to: she might remind herself of the importance of smooth property transactions, honor, or the significance of a good reputation. Presumably moral education might require teaching children why the institutions of justice support many of the goods of life that we enjoy. If, however, such support fails to convince a potential knave that she should be just, and if she feels no revulsion at the thought of injustice, then, as Hume says, she has lost a considerable motive to virtue. The virtue itself, however, might be supported by a host of connected virtues—admirable pride in one's own character (I wouldn't *want* compensation for work I haven't performed, for instance) or, on the other hand, respect for one's own limitations (Who I am to judge that this is a case where I can get by with injustice without hurting anyone?)[39] Neither the natural motives associated with such virtues nor the sense of moral obligation to having a virtue conflict with Hume's view that the will always aims at goods, although they could in some cases conflict with Darwall's view that Hume intends to put forward an egoistic theory of these goods. As I have argued, however, there are convincing reasons to reject this interpretation of Hume's theory of the will.

This explanation of the motives of the just also seems consistent with Hume's claim in "Of the Original Contract," that "uniquely with the artificial virtues and

justice in particular, virtuous actions are 'performed entirely from a sense of obligation.'"[40] Darwall finds this claim puzzling, but it becomes less so when read in context: these kinds of moral duties, Hume says, "are performed entirely from a sense of obligation, *when we consider the necessities of human society, and the impossibility of supporting it, if these duties were neglected*" (emphasis added). Here Hume seems to admit the possibility of complex supporting motives such as I have suggested might help the agent tempted to injustice. Given these supporting motives, and given the account I have suggested of how more standard motives to justice can develop along with the sense that justice is part of the good agent's character, we have a variety of possible motives to justice that involve desires for goods. An agent's sense that a just act is obligatory would only be inconsistent with Hume's theory of the will if, when pressed on why she performs a just act, she can only (with honesty and self-awareness) say something like "because those are the rules" or "because reason prescribes those rules." But Hume does not describe any such agent, and one is tempted to say that he might accuse such a person of ignorance or madness. Even in his strictest claims about the artificial virtues, he appeals to the goods supporting them as if they could be resources for the just person. In other places (as in the sensible knave passages) he suggests that just persons will be concerned with honor and honesty. But these are goods as well—goods that members of civilized society are obligated to pursue.[41]

VI. Conclusion

Having excluded the possibility of motivating reason, Hume must account in some other way for the complicated nature of the various pleasures and pains humans seem to exhibit. His account is developmental and social: we have a natural tendency towards moral evaluation, and an increase in the size of society tends to increase the complexity of those evaluations. Nature does not leave our moral development to our *own* sentiments. We are educated by others and depend on them in such ways that their reactions to us affect us intimately. In order to converse with one another, we must develop moral language from the general point of view, and our natural tendency to adopt others' sentiments to our own aids in the process by which others' disapprobation becomes a reason for us to become one type of person or another. Were it not for this intricate social network, we would neither have need of nor benefit from many of the motives we in fact have in the concrete circumstances of life. The obligation or bindingness we find in morals comes not from reason, but from the interaction between our passions and society and our relationships with other people. Darwall dismisses this idea as essentially an unfortunate inheritance from Hutcheson. "With no evident philosophical rationale," he says, "Hutcheson terms moral sense's approbation itself an obligation, and Hume simply follows his lead."[42]

In suggesting that Hume transcends such a theory, it may seem that Darwall's interpretation of Hume is charitable. Darwall says that "the richness and complexity of [Hume's] thought outrun his usual categories, giving him a much more interesting view and a distinctive place in early modern thinking about obligation."[43] Here is where Hume stops trying to "explain normativity away," where he could no longer "be pleased to accept naturalist substitutes for the normative notions to which he helps himself."[44] Darwall ascribes to Hume insights about moral motivation not fully articulated until Kant's critical period. He attributes to him the intellectual virtue of not letting his system overtake his genius; his true insights of justice do not fall victim to his hedonistic, egoistic theory of the will. And he saves him from his own claim that there is sophistry in the neighborhood of justice.

But to accept this solution, we must also accept that Hume makes at least two serious errors. These are not errors in his observations about human nature but errors internal to the structure of his theory. Darwall, argues, in effect, that Hume implicitly contradicts his own theory of the will by implying that the rules themselves have normative force for agents. This theory of the will is in turn inextricably tied to Hume's centrally important conception of reason as a purely *passive* faculty that "can never be a motive to any action of the will" and "can never oppose passion in the direction of the will" (T 2.3.3.1; SBN 413). Moreover, Darwall holds Hume to be "confused" when he speaks in the *Treatise* as if the sense of obligation that is the unique motive to justice were *moral* obligation rather than rule obligation.

Fortunately for Hume, by focusing on the social and developmental aspects of his theory of justice, we can understand how it might be possible for conventions, combined with the normal mechanisms of praise and blame, to create the possibility for new kinds of motives and new character traits. Once society has thus progressed, its members can take pleasure in just acts and be pained by unjust ones in ways analogous to their pleasures and pains associated with other virtues and vices, through the normal mechanisms of moral obligation. The alternative is to rob Hume of much of his power as a representative not only of the empirical naturalist tradition, but of the tradition of ethical theory that evolved from his own work and continues to be compelling to many. If the interpretive choice is between trusting Hume's intelligence and mastery of philosophical skills, on the one hand, and claiming that he contradicts his stated premises with his own conclusions, on the other, then perhaps the truly charitable course is the former, even if it requires accepting that Hume deeply disagrees with a picture of human agency that some philosophers have come to believe to be important centuries later.

NOTES

I am indebted to many for comments and suggestions on earlier drafts of this paper, including the editors at *Hume Studies*, two anonymous readers, Stephen Darwall, members of the ethics reading group at the University of Notre Dame, David Solomon, Karl Ameriks, Robert Roberts, Thomas Hibbs, Robert Krushwitz, Michael Beaty, and other members of the philosophy department at Baylor University.

1 *The British Moralists and the Internal 'Ought'* (Cambridge: Cambridge University Press, 1995). Darwall presents his argument in a 1993 article ("Motive and Obligation in Hume's Ethics," *Nous*, 27 [1993]: 415–48) but it reappears substantially unrevised in the *British Moralists and the Internal 'Ought.'*

2 References to Hume's major works will be made parenthetically in the text as follows:

David Hume, *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton, Oxford Philosophical Texts (Oxford: Oxford University Press, 2000). Quotations from the *Treatise* will be identified as 'T' with section and paragraph numbers which will be followed by page references to David Hume, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, 2nd ed., revised by P. H. Nidditch (Oxford: Clarendon Press, 1978), identified as 'SBN.'

David Hume, *An Enquiry Concerning the Principles of Morals*, ed. Tom L. Beauchamp (Oxford: Clarendon Press, 1998). Quotations from the second *Enquiry* will be identified as 'EPM' with section and paragraph numbers which will be followed by page references to David Hume, *Enquiries concerning Human Understanding and concerning the Principles of Morals*, ed. L. A. Selby-Bigge, 3rd ed., revised by P. H. Nidditch (Oxford: Clarendon Press, 1975), identified as 'SBN.'

3 Despite these statements, there is still understandable disagreement about Hume's understanding of good, evil, pleasure, and pain. For example, Jacqueline Taylor emphasizes the socially constructive aspect of Hume's account. She says in her "Justice and the Foundations of Social Morality in Hume's *Treatise*," "I want primarily to undermine what has come to be a fairly standard view that understands Hume to be equating good and evil straightforwardly with pleasure and pain (as sensations). Hume suggests the equation himself several times . . .; but he is clearly attempting to articulate an account of the passions that underscores the importance of social relations and other social factors to our understanding of the causes, objects, meaning, and so forth of our various passionate responses" (*Hume Studies* 24 [1998]: 28, note 20). I agree with this point, but to dismiss Hume's own equivalences might suggest that Hume has an overly simplistic view of pleasure (as viciously homogenous, for example), and perhaps of goods as well. See below.

4 See *The British Moralists and the Internal 'Ought,'* 14ff.

5 Ibid., 14–15.

6 Although it may help us desire new ends. Cf. Darwall, 15–16 on the "calm reflective deliberation view."

7 Ibid., 16–17.

8 Ibid., 290.

9 Darwall's careful analysis of the history shows that sorting the early moderns is no simple task either, however. Witness Locke's inclusion in both categories.

10 Ibid., 284. In his "Concluding Reflections," Darwall describes Hume's "sceptical solution" to the problem of normativity and notes its continuing influence:

On the one hand, [such solutions] assert that there is no such thing as normativity as traditionally conceived, but they argue, on the other, that many of the traditional conception's functions are served just as well by a sufficiently close substitute that can be based in a philosophically respectable way–normativity as instrumental rationality or as calm, reflective deliberation.

There is no overestimating the power of this philosophical program, whose wide and deep influence continues to the present day. Still, even Hume feels the attraction of a more robust notion of normativity than any that can be supplied by a reductionist or reforming naturalism–at least, by those of his time (*The British Moralists*, 320–1).

11 Ibid., 287.

12 Ibid., 294.

13 Ibid., 295.

14 Ibid.

15 *Dissertation on the Passions*, in David Hume, *The Philosophical Works*, ed. T. H. Green and T. H. Grose, 4 vols. (London: Longman, Green, 1898) 4: 139.

16 *The Invention of Autonomy: A History of Modern Moral Philosophy* (Cambridge: Cambridge University Press, 1998), 371.

17 See T 3.2.1.8; SBN 479 and T 3.2.5.12; SBN 523.

18 Darwall argues that rule obligation is the only way to make sense of Hume's remark that "After this convention, concerning abstinence from the possessions of others, is enter'd into, and every one has acquir'd a stability in his possessions, there immediately arise the ideas of justice and injustice; as also those of *property*, *right*, and *obligation*. The latter are altogether unintelligible without first understanding the former" (T 3.2.2.11; SBN 490–1). Since neither moral nor natural obligation depend on justice or conventions, "[Hume] must have some third concept in mind. And we can see pretty clearly what it must be; he must be referring to a notion of obligation defined through the rules of justice (or their acceptance) themselves" (*The British Moralists*, 297). I am not convinced that Hume's remark about the intelligibility of obligation warrants the introduction of rule obligation, nor that Humean natural obligation is best understood as always equivalent to self-interest, but these issues are beyond the scope of this paper.

19 *The British Moralists*, 297.

20 What Hume means by "first" here is as yet ambiguous. I will address this question below, but (1) and (2) remain distinct regardless of the answer to this interpretative question. We may be assured of this distinction in a way that Hume could not have been. Kant clarifies the issue by accepting a version of (1)—insisting that the intention behind an action is the sole basis on which we may proclaim an act right or wrong—but denying

(2)—arguing that the perception that an act is the right one to perform is *the* proper motive to morally worthy action. Of course, as Kant and Hume disagree on the possibilities for varieties of motives, they would interpret these claims in different ways.

21 See also Hume's letter to Francis Hutcheson of 17 Sept. 1739: "Actions are not virtuous nor vicious; but only so far as they are proofs of certain Qualitys or durable Principles in the Mind" (*The Letters of David Hume*, ed. J. Y. T. Greig, 2 vols. [Oxford: Clarendon Press, 1932]), 1:34, and the related discussion of "virtue in rags" in "Of the origin of the natural virtues and vices" (T 3.3.1.19ff.; SBN 584ff.).

22 *The British Moralists*, 301. See T 3.2.2.22; SBN 497 for Hume's claim that the connection between justice and interest is "somewhat singular." See also Hume's remark in his discussion of how society comes to establish justice conventions: "'[T]is only upon the supposition, that others are to imitate my example, that I can be induc'd to embrace that virtue; since nothing but this combination can render justice advantageous, or afford me any motives to conform my self to its rules" (T 3.2.2.22; SBN 498).

23 *The British Moralists*, 299.

24 Interestingly, Hume claims that humanity's solution to this problem is the establishment of government, to "render the observance of justice the immediate interest of some particular persons, and its violation their more remote" (T 3.2.7.6; SBN 537). Darwall does not discuss this solution, but I suspect it must be taken into account in any complete explication of Hume's views on justice.

25 *The British Moralists*, 300.

26 I am suggesting, of course, that Hume recognizes something like the free-rider problem much sooner than Darwall believes that he does, although Hume probably would approach this problem from a somewhat different perspective. A thorough discussion of this question is beyond the scope of this paper. The material point here is that interest will only lead to compliance if enough comply, and Hume recognizes and grapples with the continual tendency towards dangerous levels of non-compliance. See also note 32.

27 Darwall cites Hume's claim that "we have naturally no real or universal motive for observing the laws of equity, but the very equity and merit of that observance" (T 3.2.1.17; SBN 483).

28 He also considers briefly a tendency to blow the disadvantages of injustice out of proportion in such a way that one always believes that just acts will be in one's long-term interest (316). But Darwall rightly concludes that Hume would have no patience with such a trait.

29 See T 2.3.3.10; SBN 418 for a description of strength of mind in terms of calm and violent passions. Note also EPM 4.1; SBN 205: "Had every man sufficient *sagacity* to perceive, at all times, the strong interest which binds him to the observance of justice and equity, and *strength of mind* sufficient to persevere in a steady adherence to a general and distant interest, in opposition to the allurements of present pleasure and advantage; there had never, in that case, been any such thing as government or political society, but each man, following his natural liberty, had lived in entire peace and harmony with all others. What need of positive law where natural justice is, of itself, a sufficient restraint?"

30 *The British Moralists*, 309.

31 Ibid., 307.

32 Darwall does not explain the problem exactly this way, but perhaps this is the point he refers to on 311 when he says, "Adding in the moral interest in avoiding self-hatred does not help; indeed, it exacerbates the problem." He also cites this passage from "The Origin of Government" as evidence that Hume "finally, at least" recognizes the free rider problem: "All men are sensible of the necessity of justice to maintain peace and order; and all men are sensible of the necessity of peace and order for the maintenance of society. Yet notwithstanding this strong and obvious necessity, such is the frailty and perverseness of our nature! . . . Some extraordinary circumstances may happen, in which a man finds his interests to be more promoted by fraud or rapine, than hurt by the breach which his injustice makes in the social union. But much more frequently, he is seduced from his great and important, but distant interests, by the allurement of present, though often very frivolous temptations" (*Essays Moral, Political, and Literary*, ed. Eugene F. Miller [Indianapolis: Liberty Fund, 1985], 38). The concerns behind the free rider problem are not so absent from the *Treatise* as Darwall's "finally, at least" might imply, especially in the discussion of the origins of the government. It is possible, however, that Hume's different views about the nature of ethics might lead him to assign such concerns a far different importance than some contemporary philosophers do. A full discussion of these issues, however, is beyond the scope of this paper.

33 *The British Moralists*, 315.

34 See Hume's remark at T. 3.1.2.9 (SBN 475) that "Perhaps it will appear afterwards, that our sense of some virtues is artificial, and that of others natural." An anonymous reader of an earlier draft of this paper suggested that discussing the circle argument in terms of an evolution or genealogy of justice is to misunderstand its import, since the relevant passages "do not concern the origin of motives to justice" but the "question: of what motive do we approve in approving of acts as just." My suggestion here is precisely that, in the case of Hume's artificial virtues, one cannot divide these two concerns. We cannot understand the motive that we approve in approving of acts as just except in terms of the evolution of justice itself. Moreover, as I say in the body of the text, the surrounding context makes it clear that Hume is less concerned in these particular passages with describing motives of justice than with establishing that justice is an artificial virtue. If his concern here were to explain what motive we approve when we approve of just acts, we must judge this section of the text an abject failure—so much so, indeed, that Hume's "failure" provides evidence that this was not the test he was trying to pass.

35 I am grateful to the editors of *Hume Studies* for suggesting that this point be put in this way.

36 Of course, Hume offers such explanations in the discussion following his argument that justice is an artificial virtue.

37 Hume summarizes these stages at T 3.2.6.11; SBN 533: "Upon the whole, then, we are to consider this distinction betwixt justice and injustice, as having two different foundations, *viz.*, that of *self-interest*, when men observe, that 'tis impossible to live in society without restraining themselves by certain rules; and that of *morality*, when this interest is once observ'd to be common to all mankind, and men receive a pleasure from

the view of such actions as tend to the peace of society, and an uneasiness from such as are contrary to it. 'Tis the voluntary convention and artifice of men, which makes the first interest take place; and therefore those laws of justice are so far to be consider'd as *artificial*. After that interest is once establish'd and acknowledg'd, the sense of morality in the observance of these rules follows *naturally*, and of itself; tho' 'tis certain, that it is also augmented by a new *artifice*, and that the public instructions of politicians, and the private education of parents, contribute to the giving us a sense of honour and duty in the strict regulation of our actions with regard to the properties of others."

38 *The British Moralists*, 304.

39 I am indebted to Robert Roberts for the suggestion that some unity of the virtues might be required for justice in the civilized state.

40 "Of the Original Contract," in David Hume, *Essays Moral, Political, and Literary*, 480.

41 This aspect of Hume's theory saves him from accusations of utilitarianism when it comes to justice; the agent need not in any sense be attempting to maximize goods. She fully believes that honorable actions, the system of property, etc. are goods, and her respect for her own character (if she is virtuous) will not allow her to disrespect such goods.

42 "Motive and Obligation in Hume's Ethics," 421.

43 *The British Moralists*, 287.

44 Ibid., 321.