Locke and Hume both recognize an obligation to respect others’ property and an obligation to keep one’s promises, as well as a further obligation to obey one’s government; but their accounts of these three obligations differ radically. On Locke’s view, the first two obligations—to respect property and to keep promises—are original moral obligations, founded in divinely instituted natural law that exists prior to and independent of all human conventions; while the third obligation—to obey government—originates in a contract that is motivated by the desire to better protect one’s property and that derives its obligatory force from the natural obligation to keep one’s promises. On Hume’s more naturalistic view, in contrast, each of the three obligations—to what he usually calls “justice,” “fidelity,” and “allegiance,” respectively—depends on the existence of its own distinctive convention. Furthermore, while he allows that government as an historical institution is likely to have begun with an actual promise, he holds that the three species of obligations themselves are entirely distinct and originally non-moral: none derives its force as an obligation from either of the others, and each was a self-interested obligation before it was a moral obligation.

In *A Treatise of Human Nature*, Hume highlights his difference with Locke and others on the conventional origin of the obligations to justice and fidelity by calling justice and fidelity “artificial virtues”—by which he means that it is only “by means of an artifice or contrivance” that they produce the moral approbation that constitutes them as virtues (T 3.2.1.1; SBN 477). In both the case of justice (*Treatise* 3.2.1-4) and the case of fidelity (*Treatise* 3.2.5), he begins by arguing that the virtue in question is “artificial” before going on to explain how the convention from which it is derived originated and why adherence to that convention is regarded as a virtue.

Central to his defense of his anti-Lockean claims that justice and fidelity are both artificial virtues is a line of argument that has come to be known as his “Circle Argument.” Hume applies this line of argument first to justice in *Treatise* 3.2.1 (“Justice, whether a natural or artificial virtue?”) and then to fidelity in *Treatise* 3.2.5 (“Of the obligation of promises”). Both versions of the argument depend on a core thesis of his virtue-based approach to ethics: that the moral merit of an action is derived entirely from the moral merit of the virtuous motive of which it is a sign. From this core thesis, he derives the general principle—which we may call the “First Virtuous Motive Principle”—that for any virtuous action, there must be a “first virtuous motive” that is other than a sense of duty to the action itself. This principle follows from the core thesis of his virtue ethics, he argues, because the only alternative would be a vicious “circle,” in which the

---

1 Hume uses the term ‘justice’ both in a narrow and a broad sense. For while he sometimes treats and defines “justice” as concerned exclusively with the rules of property, at other times he explicitly includes “fidelity” (which concerns all promises, whether concerned with property or not) within its scope. I shall use the term ‘justice’ in the narrow sense in which it is concerned only with property and is distinct from fidelity. Hume also uses the terms ‘honesty’ and ‘equity’ in connection with the obligation to respect property.

2 References to the *Treatise* by book, part, section, and paragraph, are to Hume 2000; the page reference following ‘SBN’ is to Hume 1978a.

3 The virtue of the motive is, in turn, discerned by the moral sense, by means of sentiments of moral approbation that usually depend, in part or in whole, on sympathy with those who possess or are otherwise affected by the motive (*Treatise* 3.1.2, T 3.3.1).
moral merit of the action would have to be derived from the antecedent virtue of the motive, while the virtue of the motive could only be derived, in turn, from the antecedent moral merit of the action. Duty can function as a motive, and even (in its way) a morally praiseworthy one, on Hume’s account; but it can do so only when the moral merit of a kind of action has already been established by its relation to another virtuous motive. In acting from duty, he holds, we seek either to hide from ourselves our lack of that virtuous motive or to inculcate it in ourselves through practice (T 3.2.1.8; SBN 479). Once he has established the First Virtuous Motive Principle, Hume completes the Circle Argument by applying it to observations about the motives to acts of justice and fidelity, respectively, in order to argue that—on pain of circularity—each must be an artificial virtue.

Hume’s use of this line of argument has seemed notoriously problematic, however, for three main reasons. First, it appears that, despite the requirements of his own First Virtuous Motive Principle, Hume does not allow that there are any virtuous motives that can explain the full range of acts of justice or fidelity other than the sense of duty to perform them. On the contrary, there are several passages in which he seems explicitly to deny that there are any such non-moral motives (whether themselves virtuous or not) sufficient to motivate justice, fidelity, or both.

Second, it appears that, given his survey of possible motives, Hume should not allow that there is any first virtuous motive for justice or of fidelity other than the sense of duty. He argues that two of the most obvious contenders—public benevolence and private benevolence—are not themselves sufficient to motivate the full range of just actions, and he also notes that benevolence and gratitude are not themselves sufficient to motivate all acts of fidelity. Although he does clearly grant a primary role to “self-interest” in the origins of both justice and fidelity, he also seems to allow that, like public and private benevolence, it, too, cannot ultimately motivate the full range of acts of justice or fidelity. Furthermore, and independent of its scope, self-interest does not appear to be a virtuous or morally praiseworthy motive at all.

Third, it appears that Hume cannot allow that there are “first virtuous” non-moral motives to justice and fidelity without contradicting his own conative psychology. For he seems to allow, and even to insist, that consistent adherence to the requirements of justice or fidelity would require that one apply rules to regulate one’s conduct independently of considerations of pleasure and pain, thereby taking those rules as authoritative. Yet his theory of motivation seems to require that only present or prospective pleasure or pain can ultimately motivate action, and it seems to allow no way to regard a rule as authoritative.

All of this seems quite disastrous for Hume. For since he holds that the First Virtuous Motive Principle is a direct consequence of the core thesis of his virtue ethics, he cannot deny that there are motives satisfying its requirements for justice and fidelity without allowing either (i) that his own virtue ethics is false or (ii) that he is wrong to characterize justice and fidelity as virtues.⁴

---

⁴ Francis Snare (1991) and Rachel Cohon (1997) have proposed that Hume should be interpreted as making the former allowance; Knud Haakonssen (1981) and David Gauthier (1992) have suggested that Hume may (ultimately, at least) be interpreted as making the latter allowance. Both of these lines of interpretation are desperate remedies, however. Contrary to the former, Hume must rely on the core thesis of his virtue ethics and the First Virtuous Motive Principle that he derives from it not only as key premises in his argument that justice is artificial but again to support his argument that fidelity is artificial, which occurs only after his discussion of justice is quite concluded and beyond revision; and indeed he never issues anything like a recantation of either the thesis or the principle. Contrary to the latter, his lists of the virtues consistently include justice and fidelity, and at no point does he show any inclination to revise his classification of them; indeed, one of the final appendices to An Enquiry concerning the Principles of Morals (his last exclusively moral work) is devoted to them specifically as virtues.
In the first section of this paper, I will consider whether Hume does allow that there are first virtuous motives to justice and fidelity other than a sense of duty. In the second, I will consider whether he should allow that there are such motives in light of his survey of the available alternatives. In the third section, I will consider whether he can consistently allow such motives within the constraints of his conative psychology. My conclusions will be that Hume can, should, and does provide such motives. Commentators have typically characterized Hume’s Circle Argument as involving irreconcilable contradictions or difficulties; but with respect to these three questions, at least, Hume’s use of the Circle Argument can be squared.

I. Does Hume Allow First Virtuous Non-Moral Motives to Justice and Fidelity?

In a seminal article, David Gauthier (1992) emphasized a passage in the Treatise in which Hume appears to deny that there is any non-moral motive at all to justice, and another in which Hume appears to deny that there is any non-moral motive to fidelity. Both passages have subsequently been much-cited. In an important subsequent article, Stephen Darwall (1993) has emphasized a further passage from a later essay, “Of the Original Contract,” in which Hume appears to deny that there is any non-moral motive for any artificial virtue. In order to understand Gauthier’s two passages, it will be necessary to examine the versions of the Circle Argument in which they are embedded. In order to understand Darwall’s passage, it will be necessary to examine Hume’s theory of obligation.

Does Hume Reject Non-Moral Motives to Justice? The passage cited by Gauthier occurs near the conclusion of his application of the Circle Argument to the case of justice:

[W]e have naturally no real or universal motive for observing the laws of equity [i.e., the rules of justice], 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. (T 3.2.1.17; SBN 483)

It should be noted at the outset that the word ‘naturally’ does not appear in the first-edition version of this passage; Hume added it to his own printed copy of the Treatise, presumably as a correction intended for a possible second edition. In order to appreciate the significance of its addition and its relation to the argument as a whole, it is necessary to note that, just prior to undertaking the discussion of justice, Hume distinguishes five senses of the term ‘natural’ by means of relevant contrasts: what is “natural” may be “opposed to miracles,” “opposed to what is rare and unusual,” “opposed to artifice,” “opposed to civil” and “opposed … to moral.” He adds that, so far as the use of the term in the remainder of the Treatise is concerned, “the opposition...
will always discover the sense, in which it is taken.” Accordingly, when quoting Hume in what follows, I try to disambiguate the term, using brackets, wherever it occurs.

Hume begins his application of the Circle Argument to justice with two observations about moral evaluative practice:

J1. When we praise any actions, we regard only the motives that produced them, and consider the actions as signs or indications of certain principles in the mind and temper.

J2. When we require any action, or blame a person for not performing it, we always suppose, that one in that situation shou’d be influenc’d by the proper motive of that action, and we esteem it vicious in him to be regardless of it.

He uses these observations to support the core thesis of his virtue ethics:

J3. All virtuous actions derive their merit only from virtuous motives, and are consider’d merely as signs of those motives. (from J1 and J2)

In the piece of reasoning that gives the Circle Argument its name, he then makes what he takes to be an uncontroversial logical point and applies it to his virtue ethics principle to deduce the First Virtuous Motive Principle:

J4. An action must be virtuous before we can have a regard to its virtue.

J5. 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. (from J4)

J6. [First Virtuous Motive Principle] The first virtuous motive, which bestows a merit on any action, can never be a regard to the virtue of that action, but must be some other natural [i.e., non-moral] motive or principle. (from J3 and J5)

From this principle, Hume infers what he calls “an undoubted maxim”:

J7. [Undoubted Maxim] 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. (from J6)

Commentators have generally not distinguished this Undoubted Maxim (as I shall follow Hume in calling it) from the First Virtuous Motive Principle from which it is derived, but the difference between them will prove to be important.

Suggesting that there will be “great difficulty” in finding a motive to just actions that will satisfy the Undoubted Maxim, he surveys three contenders—self-interest (i.e., private interest or self-love), public benevolence, and private benevolence:

---

6 In his primary discussions of justice and fidelity (Treatise 3.2.1-5), he uses the term ‘natural’ in three of these five senses—occasionally to mean “not unusual” and sometimes to mean “non-moral,” but usually to mean “non-artificial”—often relying on context alone, as promised, to indicate the intended meaning.

7 This statement is sometimes read as claiming that the “first virtuous motive” to any action must be some other non-artificial motive, rather than some other non-moral motive. Such a reading is implausible, however, for Hume has at this point provided no premises about artificial or non-artificial motives from which such a conclusion could be drawn. Moreover, a restriction to non-artificial motives would have no point, since step J6 is used only to derive step J7, which concerns non-moral motives. Hume’s use of ‘other’ in ‘some other natural motive’ suggests that he is contrasting the obviously moral character of “a regard to the virtue of the action” with the non-moral character of the “first virtuous motive.”
J8. Should we say, that a concern for our private interest or reputation, is the legitimate motive to all honest [i.e., just] actions: it would follow that wherever that concern ceases, honesty can no longer have place.

J9. Self-love, when it acts at its liberty, instead of engaging us to honest [i.e., just] actions, is the source of all injustice and violence; nor can a man ever correct those vices, without correcting and restraining the natural movements [i.e., movements prior to artifice and contrivance] of that appetite.

J10. Public benevolence … or a regard to the interests of mankind, cannot be the original motive to justice.  

J11. Private benevolence, or a regard to the interests of the party concern’d [cannot be] the original motive to justice.

From this survey of possible original motives, he concludes:

J12. We have naturally [i.e., non-artificially] no real or universal motive for observing the laws of equity [i.e., of justice] but the very equity and merit of that observance. (from J8, J9, J10, and J11)

Step J12 is, of course, the first clause of the passage cited by Gauthier; the remainder of that passage simply re-iterates the threat of the circularity, first identified in step J5, that would result from the Undoubted Maxim of step J7 if there were no original non-moral motive to justice at all. On this basis, Hume draws his final conclusion about the artificiality of justice:

J13. Unless we will allow that nature has established a sophistry, and rendered it necessary and unavoidable, we must allow, that the sense of justice and injustice is not derived from nature, but arises artificially, though necessarily, from education and human conventions. (from J5, J7, and J12)

Gauthier fully allows that Step J12 may be read as claiming only that we have no non-artificial motive to justice; but in fact, it can only be read in that fashion. Not only is that the only reading

8 Hume provides three premises in support of this claim:
   (i) Public interest is not naturally [i.e., non-artificially] attach’d to the observation of the rules of justice, but is only connected with it, after an artificial convention for the establishment of these rules.
   (ii) If we suppose, that [a] loan was secret, and that it is necessary for the interest of the person, that the money be restor’d in the same manner (as when the lender wou’d conceal his riches), in that case the example ceases, and the public is no longer interested in the actions of the borrower; tho’ … there is no moralist, who will affirm, that the duty and obligation ceases.
   (iii) [M]en, in the ordinary conduct of life, look not so far as the public interest, when they … abstain from theft, and robbery, and injustice of every kind.

9 Hume provides three premises in support of this claim:
   (i) [If] the party concern’d … be my enemy, and has given me just cause to hate him … [or] he be a vicious man, and deserves the hatred of all mankind … [or] he be a miser, and can make no use of what I wou’d deprive him of … [or] he be a profligate debauchee, and wou’d rather receive harm than benefit from large possessions … [or] or I be in necessity, and have urgent motives to acquire something to my family … in all these cases, the original motive to justice wou’d fail.
   (ii) Were private benevolence the original motive to justice, a man wou’d not be oblig’d to leave others in possession of more than he is oblig’d to give them [or] at least the difference wou’d be very inconsiderable.
   (iii) A man’s property is suppos’d to be fenc’d against every mortal, in every possible case. But private benevolence towards the proprietor is, and ought to be, weaker in some persons, than in others: And in many, or indeed in most persons, must absolutely fail.
that makes sense of step J13’s conclusion that the sense of justice is (barring an impossible “sophistry in nature”) artificial, but it is also the only reading that allows step J12 to follow from steps J8-J11, given step J8’s explicit limitation to the “natural” (i.e., non-artificial) movements of self-interest or self-love. Moreover, to read the passage as denying that there is any motive to justice other than duty would conflict directly with Hume’s own subsequent claim that “self-love [i.e., self-interest]… produces the rules of justice, and is the first motive of their observance” (T 3.2.8.5; SBN 543).

**Does Hume Reject Non-Moral Motives to Fidelity?** Gauthier also called attention to a parallel passage in the *Treatise* in which Hume appears to deny that there is any motive to fidelity other than the sense of duty.

Now, ’tis evident we have no motive leading us to the performance of promises, distinct from a sense of duty. If we thought, that promises had no moral obligation, we never shou’d feel any inclination to observe them. (T 3.2.5.6; SBN 518-19)

This passage occurs near the conclusion of Hume’s extension of the Circle Argument to fidelity. Hume begins the section “Of the obligation of promises” (T 3.2.5.1; SBN 516) by describing his strategy for demonstrating the artificiality of fidelity:

That the rule of morality, which enjoins the performance of promises, is not natural [i.e., non-artificial], will sufficiently appear from these two propositions, which I proceed to prove, viz. that a promise would not be intelligible before human conventions had established it; and that even if it were intelligible [before human conventions had established it], it would not be attended with any moral obligation. (T 3.2.5.1; SBN 516)

Throughout the execution of this strategy, Hume emphasizes the limitation of the circumstances he is considering to “natural” ones, not involving human conventions. Hume’s extension of the Circle Argument to fidelity occurs at the conclusion of this strategy. He introduces it by remarking that “the same truth” just established—i.e., that “there could not naturally [i.e., non-artificially] … arise any obligation from a promise” (T 3.2.5.5; SBN 518)—“may be prov’d still

---

10 He argues for the first proposition—i.e., that a promise would not be intelligible before human conventions had established it—by surveying various acts of the mind that do not involve conventions and showing that none of them is equivalent to promising; these acts include resolving, desiring, and willing the promised action. This result suggests that if promising is to be any natural act of the mind, it must somehow be, not the willing of the promised action, but the willing of the obligation itself. Yet obligations result, Hume claims, only from agreeable or disagreeable sentiments with respect to an action, and we cannot “naturally” (i.e., directly, without some artifice and contrivance) change our sentiments by an act of will. Hume later refers back to this conclusion—in a passage that Gauthier also cites—when he remarks that “we have prov’d already, that there is no such act of the mind [as naturally willing an obligation in a promise], and consequently that promises impose no natural obligation” (T 3.2.5.12; SBN 523). There is no violation of the First Virtuous Motive Principle in this passage, since ‘natural obligation’ again means “non-artificial obligation,” and an account of the obligation to fidelity as artificial does not require that promising be an act of willing an obligation.

Hume then offers two arguments for his second proposition—i.e., that even if promises were intelligible prior to human conventions, such pre-conventional “promises” could not “naturally [i.e., non-artificially] be attended with any moral obligation” (T 3.2.5.5; SBN 518). The first argument for this proposition appeals to the same set of considerations that ruled out the suggestion that promising could be the natural willing of an obligation.

This appears evidently from the foregoing reasoning. A promise creates a new obligation. A new obligation supposes new sentiments to arise. The will never creates new sentiments. There could not naturally [i.e., non-artificially], therefore, arise any obligation from a promise, even supposing the mind could fall into the absurdity of willing that obligation. (T 3.2.5.5; SBN 518)
more evidently by that reasoning, which prov’d justice in general to be an artificial virtue” (T 3.2.5.6; SBN 518). Thus, by the time he comes to apply the Circle Argument to fidelity, he can assume that the restriction of his negative arguments to circumstances that are “natural,” in the sense of non-artificial, is well-understood, and he makes it explicit only intermittently—though still often—in the course of the argument itself.\footnote{Baldwin 2004 also mention, more generally, that the context of the argument suggests a limitation of the passage Gauthier cites to natural (i.e., non-artificial) motives. Gauthier allows that this interpretation of the passage is possible.}

Hume begins this version of the Circle Argument with a formulation of the Undoubted Maxim that he has already derived from his virtue ethics by means of the First Virtuous Motive Principle in the case of justice:

F1. No action can be requir’d of us as our duty, unless there be implanted in human nature some actuating passion or motive, capable of producing that action. (from J3)

He then offers an analogue of his earlier logical point, phrased this time in terms of duty and obligation:

F2. A sense of duty supposes an antecedent obligation. (from J4)

Next, he gives a three-step argument to show that a “natural” obligation to an action requires a “natural” passion to motivate it:

F3. Where an action is not requir’d by any natural [i.e., non-artificial] passion … it may be omitted without proving any [non-artificial] defect or imperfection in the mind or temper.

F4. Where an action is not requir’d by any natural [i.e., non-artificial] passion it may be omitted without proving any … [non-artificial] vice. (from F3)

F5. Where an action is not requir’d by any natural [i.e., non-artificial] passion it cannot be required by any natural [i.e., non-artificial] obligation. (from F4)\footnote{The line of argument in steps F3-F5 parallels a line of argument that Hume has already offered—also limited, implicitly but clearly, to non-artificial circumstances (i.e., “in uncultivated nature”)—with respect to justice (Treatise 3.2.2.8; SBN 488).}

In this argument, the unstated limitation of steps F3 and F4 to non-artificial defects and vices (i.e., to conditions that could be defects or vices independent of artifice) is required in two ways. First, it is required for plausibility. For Hume allows that conventions can give rise to “new motives”\footnote{“They are the conventions of men, which create a new motive, when experience has taught us that human affairs would be conducted much more for mutual advantage, were there certain symbols or signs instituted, by which we might give each other security of our conduct in any particular incident. After these signs are instituted, whoever uses them is immediately bound by his interest to execute his engagements, and must never expect to be trusted any more if he refuse to perform what he promised.” (Treatise 3.2.5.10; SBN 522)} or passions—such as desires concerning the rules of those conventions—and he has done nothing to argue that the lack of such passions, in the presence of a convention, could not constitute a new defect or vice. Second, the limitation is required in order to explain the explicit limitation to “natural” (i.e., non-artificial) obligations in the conclusion at step F5.

From this conclusion and the previous logical point, he infers that it must be some motive other than duty itself that makes possible a natural (i.e., non-artificial) duty:
F6. The motive [implanted in human nature through which an action can be requir’d of us as our (non-artificial) duty] cannot be the sense of duty. (F2 and F5)

The restriction in the scope of step F6 to non-artificial duties is required by the explicit restriction to *natural* obligations in step F5, from which it is derived.

Next, Hume observes:

F7. If we thought, that promises had no moral obligation, we shou’d never feel any [non-artificial] inclination to observe them.

From this, he infers:

F8. We have no [non-artificial] motive leading us to the performance of promises, distinct from a sense of duty. (from F7)

And from this, in turn, he infers:

F9. There is naturally [i.e., non-artificially] no inclination to observe promises, distinct from a sense of their obligation. (from F8)

Finally, from this plus his initial virtue-ethics principle, and his elimination of duty itself as a possible original motive to acts required by natural duties, he concludes:

F10. Fidelity is no natural [i.e., no non-artificial] virtue, and ... promises have no force, antecedent to human conventions. (F1, F6, and F9)

Steps F7 and F8 together constitute Gauthier’s second passage. Their implicit restriction to non-artificial motives and inclinations is justified by the general and often-expressed restriction of the entire argument (and of the larger strategy of which it is a part) to the case of non-artificial circumstances; but it is also needed in order to make sense of the explicit restrictions to “natural inclinations,” “natural virtues,” and circumstances “antecedent to human conventions” that occur in steps F9 and F10, which are derived from Steps F7 and F8. Thus, Hume is not denying that there is any motive to fidelity other than a sense of duty, but only that there is no such motive independent of artifice and convention. On the contrary, he immediately goes on to argue that self-interest leads individuals to adopt a convention through which they do have a “new motive” to keep their promises, concluding that “interest is the first obligation to the performance of promises” (T 3.2.5.15; SBN 522).

*Does Hume Reject Non-Moral Motives for Artificial Virtues?* The passage cited by Darwall occurs in the essay “Of the Original Contract,” where Hume distinguishes two classes of moral duties corresponding to the natural and artificial virtues, respectively:

All moral duties may be divided into two kinds: The first are those, to which men are impelled by a natural instinct or immediate propensity, which operates on them, independent of all ideas of obligation ….

The second kind of moral duties are such as are not supported by any original instinct of nature, but are performed entirely from a sense of obligation. (Hume 1978b, “Of the Original Contract”: 479-80)

If the “sense of obligation” which “entirely” motivates virtuous acts of this second kind is an exclusively *moral* sense of obligation, as Darwall asserts it is (1993: 438), then these actions would clearly violate the First Virtuous Motive Principle.
In fact, however, Hume does not say that these virtuous acts are performed entirely from a sense of moral obligation, and he gives every indication that they are not. In the practical realm, he regularly distinguishes and contrasts two species of obligation: (i) moral obligation and (ii) “the natural [i.e., non-moral] obligation of interest” (T 3.2.9.3; SBN 551; and T 3.2.11.4; SBN 556), which he also calls more simply “natural obligation” or “interested obligation.” He explains moral obligation as follows:

All morality depends upon our sentiments; and when any action or quality of the mind pleases us after a certain manner we say it is virtuous; and when the neglect or nonperformance of it displeases us after a like manner, we say that we lie under an obligation to perform it. A change of the obligation supposes a change of the sentiment; and a creation of a new obligation supposes some new sentiment to arise. (T 3.2.5.4; SBN 517)

The manner, or kind, of displeasure relevant to moral obligation is, of course, the sentiment of moral disapprobation that, along with the sentiment of moral approbation, plays a central role in his moral sense theory. He does not, in contrast, specifically explain the nature of interested obligation. If, however, moral obligation is constituted by displeasure in one certain manner at the neglect or non-performance of an action, then it is reasonable to suppose that natural or interested obligation is constituted by displeasure at the neglect or non-performance of an action in another manner—specifically, displeasure at the realization that the neglect of an act leaves one worse off than one would have been had the act been performed, so that neglect harms one’s interests.

Gauthier proposes that, while moral obligation is created by sentiments of moral approbation, “the motive associated with a redirected passion constitutes a natural obligation. The redirection curbs the natural force of the passion, restraining it so that it may be ‘better satisfy’d’” (1992: 410). Annette Baier proposes that “anything others count on or expect of us counts as an ‘obligation,’ a tie to them. So … self-interest may, if socially coordinated, generate obligations” (1992: 439). But neither of these interpretations accords with all of Hume’s examples of natural or self-interested obligation. He describes, for example “very strong motives of interest or safety to forbear any action” as, in general, constituting an “obligation” (T 2.1.10.6; SBN 312), without any mention either of motives restraining themselves for their own better satisfaction or of social coordination.

Similarly, near the end of An Enquiry concerning the Principles of Morals, Hume remarks:

Having explained the moral approbation attending merit or virtue, there remains nothing, but briefly to consider our interested obligation to it, and to enquire, whether every man, who has any regard to his own happiness and welfare, will not best find his account in the practice of every moral duty. (EPM 9.1.1; SBN 278)
One of the primary aims of the essay “Of the Original Contract” is to refute a central Lockean claim (previously mentioned) adopted by members of the Whig party—namely, that the moral duty to allegiance is derived from the moral duty to fidelity through a literal “social contract.” In the passage that Darwall cites, Hume is arguing that allegiance, like justice and fidelity but unlike the natural virtues, derives its force as a moral duty entirely from the sense of obligation that results from a consideration of what he there calls “the necessities of human society, and the impossibility of supporting it, if these duties were neglected.” But this consideration is not originally a moral consideration, as his Treatise accounts of the origins of these three virtues make especially clear. On the contrary, Hume holds that a “consideration of the necessities of human society” gives rise first to a sense of natural, or interested obligation to justice, fidelity, and allegiance. For participation in human society is, he argues, essential to the well-being of every human being; and justice, fidelity, and (ultimately) allegiance are essential to the maintenance of society. We are therefore motivated initially by an appreciation of the serious harm to ourselves that would result without the benefits of society that only justice, fidelity, and allegiance can provide us. This is then followed, on his account, by a sense of moral obligation, when sentiments arise from reflecting sympathetically on the positive effects of justice, fidelity, and allegiance—and on the negative effects of their neglect—on everyone in society.

Here he seems to treat the question of whether we have an interested obligation to act as moral duty requires as depending entirely on whether such action is most conducive to one’s own “happiness and welfare.” The implication is that we may have an interested obligation to all virtuous action—even though acting as required by natural virtues need not involve any redirected passions, on Hume’s account. And although even such self-benefiting virtues as prudence may well create social expectations, Hume’s defense of the claim that there is an interested obligation to virtue does not mention such expectations. Finally, Baier’s interpretation cannot readily explain why there seem to be only two species of obligation—moral and interested—on Hume’s account, since, given human foolishness, other people may well “count on” one regularly doing certain things that are neither morally required nor in one’s own interest. And Gauthier’s interpretation cannot readily explain why Hume regards both moral obligation and natural obligation as members of a common species, since the redirection of passions and the feeling of moral approval or disapproval have little obvious in common.

It is worth noting that not every case of preference or choice is a case in which one has a natural obligation to choose one of the alternatives; for although one might happen to enjoy one alternative more than another, it does not follow that either alternative would leave one sufficiently worse off in the aftermath that one would be displeased with oneself for having chosen it. And this corresponds to Hume’s usage, for he treats only choices in which one’s ongoing interests may be harmed as involving a natural obligation. It is, of course, quite possible that one should have both a natural and a moral obligation to perform the same action.

The passage reads in full:

The second kind of moral duties are such as are not supported by any original instinct of nature, but 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. It is thus justice or a regard to the property of others, fidelity or the observance of promises, become obligatory, and acquire an authority over mankind. For as it is evident, that every man loves himself better than any other person, he is naturally impelled to extend his acquisitions as much as possible; and nothing can restrain him in this propensity, but reflection and experience, by which he learns the pernicious effects of that licence, and the total dissolution of society which must ensue from it. His original inclination, therefore, or instinct, is here checked and restrained by a subsequent judgment or observation. The case is precisely the same with the political or civil duty of allegiance, as with the natural duties of justice and fidelity. Our primary instincts lead us, either to indulge ourselves in unlimited freedom, or to seek dominion over others: And it is reflection only, which engages us to sacrifice such strong passions to the interests of peace and public order. A small degree of experience and observation suffices to teach us, that society cannot possibly be maintained without the authority of magistrates, and that this authority must soon fall into contempt, where exact obedience is not payed to it. The observation of these general and obvious interests is the source of all allegiance, and of that moral obligation, which we attribute to it. (Essays, “Of the Original Contract”: 480)
Accordingly, in the passage that Darwall cites, Hume is claiming only that, unlike actions manifesting natural virtues, adherence to justice, fidelity, and allegiance must ultimately be motivated by a sense of obligation that is either interested, moral, or both. For lacking natural (i.e., non-artificial) inclinations to adhere to them, we must instead be motivated by a sense of either the displeasing harm to our own interests or the displeasing moral disapprobation that will result from our non-performance, or of both.

II. Should Hume Allow First Virtuous Non-Moral Motives to Justice and Fidelity?

As we have now seen, Hume does not explicitly deny that there are any motives to justice and fidelity other than the sense of duty. But can he provide motives for justice and fidelity that satisfy the First Virtuous Motive Principle, or should he conclude from his survey of available motives that no such non-moral motives exist? In the first version of the Circle Argument, as we have seen, he considers three potential motives to justice. He there provides three reasons why public benevolence cannot provide such an original motive (Step J10), and three reasons why private benevolence cannot provide it (Step J11). Both sets of reasons include cases of justice (for example, returning property in a “secret loan” or to a “profligate debauchee”) that could not be motivated by the motive in question. Thus, while not denying that some particular actions required by justice might happen to be performed out of public or private benevolence, Hume denies that either of these motives alone can explain the full range of just acts that are performed—what we might call the full “behavioral profile” of justice. Hence, he concludes that neither can be the original non-moral motive giving rise to just action that is demanded by the Undoubted Maxim. It is a further question, however, whether his survey allows that self-interest, when properly “corrected and restrained,” might be such a non-moral motive.

Self-Interest in the Origin of Justice and Fidelity. In Treatise 3.2.2 (“Of the origin of justice and property”), Hume argues that “self-love” (i.e., a regard to self-interest) is the “first motive to justice” (T 3.2.8.5; SBN 543; see also T. 3.2.2.24: SBN 498), for it alone can motivate the adoption of a convention through which individuals allow one another stable possession of material goods. A convention exists among a group of individuals, as Hume explains it, when (i) each has an interest in following a particular course of conduct but only on the condition that the others follow the same course of conduct; (ii) this common interest is mutually expressed and known among the parties; and (iii) this mutual expression and knowledge serve to produce “a suitable resolution and behaviour” (T 3.2.2.10; SBN 490). Conventions thus do not require any promise. The mutual willingness to adhere to the conventionally-agreed-upon course of conduct specified by rules of justice creates a convention that bestows “stability on the possession of … external goods, [and leaves] every one in the peaceable enjoyment of what he may acquire by his fortune and industry” (T 3.2.2.9; SBN 489). The parties to this convention each serve their own interest (which may include the interests of their close friends and relatives) by “regulating their

---

18 Such a convention is valuable, because—given the “selfishness and limited generosity” of humans and the relative “scarcity and easy change” of possession of material goods—it is essential to the maintenance of human society; and human society, in turn, is essential to satisfying human beings’ basic needs. Society is essential to human well-being, he observes, for it alone allows human beings to satisfy their needs by (i) augmenting their force through combining their powers, (ii) augmenting their abilities through the specialization permitted by a division of labor, and (iii) augmenting their security through rendering mutual aid in time of need. Human beings could not be expected to foresee these advantages prior to experience with society; but fortunately, they are naturally induced to enter into society, first by the “appetite between the sexes” and then by the “natural affection” of parents for their children.
conduct” in accordance with the “strict and inflexible rules” of property, because doing so allows them to maintain society, and in society each individual enjoys far more in the way of secure material goods than would be possible without this self-restraint. Given the recalcitrance of human passions, consistent adherence to fully worked-out conventions of justice “arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it.” Yet the convention of justice is so essential to society that it is needed to make possible the family itself, for “every parent, in order to preserve peace among his children, must establish it.”

In Treatise 3.2.5 (“Of the obligation of promises”), Hume argues that “interest is the first obligation to the performance of promises” (T 3.2.5.15; SBN 522), again through the institution of a convention, but this time one that makes possible secure exchanges of goods and services when the elements in the exchanges cannot be simultaneous. Although both conventions originate in self-interest, we soon come to approve morally those who adhere to their rules and disapprove those who do not; for we feel moral sentiments when considering with sympathy the effects of their characters on the members of society who benefit from the conventions.

The Range of Self-Interest. Yet although Hume specifies that self-interest is the original motive to justice and fidelity, commentators have argued on two main grounds that it cannot ultimately satisfy the First Virtuous Motive Principle. The first of these is that Hume appears to allow, in three different passages, that there are circumstances in which acts of justice or fidelity will not be in one’s own interest; hence, it seems, self-interest can no more explain the behavioral profile of justice and fidelity than public or private benevolence can. The first of these three passages occurs in Step J8 of the Circle Argument itself:

[J8] Should we say, that a concern for our private interest or reputation, is the legitimate motive to all honest [i.e., just] actions: it would follow that wherever that concern ceases, honesty can no longer have place. (T 3.2.1.10; SBN 480)

Some commentators (e.g., Barry Stroud 1977 and Lorraine Besser-Jones unpublished) have read this remark as an admission that self-interest cannot be the original motive to justice because it cannot produce the full behavioral profile of justice. In fact, however, the passage only states a

---

[19] As Hume explains it, human beings can mutually benefit from many exchanges of goods and services; but where the exchange cannot be simultaneous, the limited benevolence and gratitude of human beings offers no assurance that the second party will perform his or her part rather than simply take advantage of the first party’s action; and this leaves the first party no reason to act. This problem is solved by an ongoing convention to utilize a verbal formula—such as ‘I promise’—that is understood to commit one to performance on pain of being henceforth excluded from future exchanges and their benefits, a sanction that provides the self-interested motive to performance that was otherwise missing.

[20] The increasing size of societies, though beneficial, creates a further threat to justice and fidelity: the harmful effects of individual acts of injustice and infidelity on one’s own and the public interest become less certain, direct, and obvious, and so the temptations to violate the rules of property in pursuit of a present desire increase. The solution to this problem is new convention, allegiance to a government, which provides some individuals with a more particular interest in protecting and enforcing the property and contract rights of others. Meanwhile, moral approval of justice and fidelity is further encouraged by the praise of politicians, who appreciate the importance of justice to society, and by parents, who see justice as a virtue that will be useful to their children in society as well as conducive to the well-being of society as a whole. As a result, these virtues come to play an especially important role in one’s reputation, so that every one who has any regard to his character, or who intends to live on good terms with mankind, must fix an inviolable law to himself, never, by any temptation, to be induced to violate those principles which are essential to a man of probity and honour. (T 3.2.2.27; SBN 501)
condition that self-interest must meet in order to provide the original motive—namely, that no actions in the behavioral profile of justice should fall outside its motivational range. His immediately following remark, Step J9, indicates, as we have seen, only that self-interest cannot naturally (i.e., without the intervention of artifice and contrivance) meet this condition. This leaves open the possibility that self-interest can meet this condition as the result of artifice.

A second passage (also cited by Stroud and Besser-Jones) occurs after Hume has offered his account of the origins of justice:

Nor is every single act of justice, considered apart, more conducive to private interest than to public; and it is easily conceived how a man may impoverish himself by a signal instance of integrity, and have reason to wish, that, with regard to that single act, the laws of justice were for a moment suspended in the universe. (T 3.2.2.22; SBN 497)

The purpose of this remark, however, is only to explain his claim in the same paragraph that tho’ the rules of justice are establish’d merely by interest, their connexion with interest is somewhat singular, and different from what may be observ’d on other occasions.

The passage thus reaffirms that self-interest is the original motive that explains the behavioral profile of justice; and it then goes on to describe the way in which it does so:

But, however single acts of justice may be contrary, either to public or private interest, it is certain that the whole plan or scheme is highly conducive, or indeed absolutely requisite, both to the support of society, and the well-being of every individual. It is impossible to separate the good from the ill. Property must be stable, and must be fixed by general rules .... And even every individual person must find himself a gainer on balancing the account; since, without justice, society must immediately dissolve, and every one must fall into that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be supposed in society. When, therefore, men have had experience enough to observe, that, whatever may be the consequence of any single act of justice, performed by a single person, yet the whole system of actions concurred in by the whole society, is infinitely advantageous to the whole, and to every part, it is not long before justice and property take place.... Taking any single act, my justice may be pernicious in every respect; and it is only upon the supposition that others are to imitate my example, that I can be induced to embrace that virtue; since nothing but this combination can render justice advantageous, or afford me any motives to conform myself to its rules.

That is, self-interest is able to motivate the full range of just acts not by favoring the just act in every case-by-case evaluation of interest, but rather by favoring the adoption of a “scheme” or policy of adhering uniformly to the rules of justice as long as others do as well—a policy that replaces the default “policy” of trying to assess the outcomes of individual acts. The “infinite” advantages of stable property come only with “general rules” which, because of their generality, “cannot separate the good from the ill” so far as particular instances are concerned. But considerations of “good and ill” in particular circumstances do not compete with the considerations in favor of the general policy of uniform adherence, on Hume’s account; rather,

---

21 Similarly, Hume’s reference to the inability of public benevolence to “operate with any force in actions so contrary to private interest as are frequently those of justice and common honesty” (T 3.2.1.10; SBN 481) also refers implicitly to a pre-conventional consideration of the benefits of just action.
considerations of the advantages of having the general policy give each individual a motive to exclude from consideration the “good and ill” of particular instances. In short, self-interest “restrains itself,” as Hume says, by motivating individuals to adopt a policy that precludes acting on the assessment of interest in particular cases in favor of uniform adherence, thus producing the full behavioral profile of justice. It is this that benevolence alone, whether public or private, cannot do.

This is by no means the end of the matter, however. For Hume remarks in a famous passage in the concluding section of An Enquiry concerning Human Understanding:

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 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. (EPM 9.2.22; SBN 282)

This passage (emphasized by Marcia Baron (1982), Gauthier, Darwall, and many others) raises clearly for the first time the threatening prospect of a different policy, which we may call the “knabish” policy: act in accordance with the rules of justice and fidelity except when one can violate the rules to one’s own considerable advantage without detection. If self-interest motivates this policy rather than the policy of uniform adherence, then it appears that self-interest cannot, after all, produce the full behavioral profiles of justice and fidelity.

In assessing the implications of this passage, however, it is important to understand what Hume’s search for an original motive requires and what it does not. If self-interest is to be the original non-moral motive to justice and fidelity, it must in fact be capable of causing human beings to will to engage in the full range of actions contained in their behavioral profiles and to do so prior to any sense of duty, so that those actions can be signs of some motive that will be judged virtuous. It is not required that every individual act of justice and fidelity should actually maximize the interests of the person who performs it (even including its positive effects on the stability of the relevant convention); nor is it required that all individuals judge that every individual act of justice and fidelity will maximize their interests. For a policy may still be—and be judged—the most advantageous policy to adopt even if it is rightly judged that there are some cases in which a violation of it would be more advantageous than adherence. Specifically, it may remain the best policy to adopt as long as there is no other feasible policy that will, when adopted, provide the benefits of those violations without also producing greater harm (for example, but not limited to, harms from allowing too many or undesired exceptions). Indeed, in order for self-interest to be the original motive to justice and fidelity, it is not even required that

22 References to An Enquiry concerning Human Understanding by book, section, and paragraph, are to Hume 1998; the page reference following ‘SBN’ is to Hume 1975.
23 Although the passage begins by mentioning only justice, Hume is here using the term ‘justice’ in the broad sense, and the remainder of the passage makes clear.
uniform adherence to their rules be the best policy to adopt, from the standpoint of self-interest, so long as no other policy actually under consideration appears to be a better policy to adopt.

Individuals faced for the first time with the question of whether to adopt a policy with respect to property and promising-keeping are very unlikely to adopt the knavish policy for two reasons. First, the policy is unlikely to occur to them. Such individuals will usually be socially inexperienced and lacking in cunning; indeed, since initiation into the conventions of justice and fidelity typically occurs in the family, such individuals will typically be children. And for obvious self-interested reasons, if for no other, no one is likely to suggest the knavish policy to them. Second, they will be in circumstances in which adopting (i.e., attempting to implement) the knavish policy would not, in fact, be or appear to be superior to uniform adherence from the point of view of self-interest. The parties to the convention will initially constitute a very small society, consisting mostly of other individuals seen every day; and the opportunities for successful violations will be correspondingly very few, while the negative consequences for discovered violations may be great. These negative consequences may include hatred and loss of material benefits to violators if their identity is discovered, and harm to the potentially fragile but essential convention of justice even when the occurrence of a violation is discovered without the identity of the violator. In addition, these individuals will typically be bad judges of opportunities for successful violation; and they will lack skill in concealment. (Although small children naturally experiment with thievery and cheating, for example, they are also—fortunately for the cause of virtue—outlandishly bad at it.) Thus, when such individuals do violate the rules—more likely from temptation than from any considered knavish policy—they will experience what Hume calls “repeated experience of the inconveniences of transgressing” (T 3.2.2.10: SBN 490) them, even in cases where successful violation seemed possible. For all of these reasons, the knavish policy will neither be nor seem preferable to such individuals. Accordingly, the possibility of that policy need not prevent self-interest from motivating adoption of universal adherence to justice and fidelity.

Self-interest can thus constitute the original motive to justice and fidelity even if later developments offer better opportunities both for the formulation of the knavish policy and for the undetected rule-violations that it recommends. For once individuals motivated by self-interest have regulated their conduct by means of the rules of justice and fidelity, prior to a sense of duty, the requirement of an original non-moral motive has already been met, regardless of the effects of later developments. In fact, however, there are also many reasons in Hume why universal adherence to justice and fidelity generally remains stable even in the face of such opportunities. First, the full determination to regulate one’s conduct by these rules is, ipso facto, a commitment to try to avoid allowing contrary considerations to weigh in one’s deliberations in cases covered by the rules; hence later alterations in circumstances may not be salient. Second, the performance of any kind of action of moderate difficulty that becomes habitual—including, presumably, adherence to these rules—may become pleasurable through the mechanism of “facility” (T 2.3.5.1-5; SBN 422-24). Third, the more one participates in society, the more likely one is to take greater pleasure, through sympathy, with all those who do or would benefit from one’s own acts of justice and fidelity. Fourth, as societies become larger, governments are instituted by convention to provide additional sanctions. Fifth, as Hume remarks in considering the “sensible knave,” the psychological and practical difficulties of trying to implement a knavish policy without detection and the potentially catastrophic consequences of failure may still render it an inferior policy to adopt even from the perspective of a narrowly material conception of self-
Sixth, as he remarks in the same passage, the awareness of such catastrophes for violators can be a source of “satisfaction” to those who do adhere uniformly to the rules. Seventh, what he calls a “reputation” for justice and fidelity—best obtainable by revealing that one has acted justly or with fidelity when one might instead have violated the rules in secrecy—may be more valuable from the point of view of long-term self-interest, and also more immediately enjoyable through sympathy, than the material benefits of violating the rules. This will be especially true once *moral approbation* for justice and fidelity arises—as it must almost certainly do soon after conventions themselves. Finally, however, and most importantly, once *moral approbation* for existing justice and fidelity does arise (and is then encouraged by “politicians and parents”), a new source of “interested obligation” to justice and fidelity is created in the “inward peace of mind, consciousness of integrity, [and] satisfactory review of our own conduct” (EPM 9.2.25: SBN 283) that comes with moral approbation of, and pride in, one’s own virtuous character. This steady approbation and pride, and the avoidance of the corresponding disapprobation and humility, is far more important to long-term happiness, Hume claims, than are mere material goods.  

*The Moral Merit of Self-Interest.* Thus, self-interest can and does, on Hume’s view, constitute the original non-moral motive to justice and fidelity. It may still be objected, however, that self-interest cannot be the “first virtuous motive, which bestows a merit on” acts of justice and fidelity, as required by the First Virtuous Motive Principle. Unrestrained self-interest, at least, is not a virtue at all—on, the contrary, it is, as Hume observes, the primary source of violence and injustice. And he seems at best diffident about whether self-interest in general may be considered a virtue.

Because he regards unrestrained self-interest as the primary cause of violence and injustice, it is often proposed that the first virtuous motive bestowing merit on acts of justice and fidelity is, for Hume, “enlightened self-interest”—meaning, presumably, a combination of self-interest with accurate knowledge of how the conventions bear on one’s own interests. Yet it seems implausible to suppose that even this could be the “first virtuous motive” responsible for the merit of acts of justice and fidelity. As Rachel Cohen has argued (1997), someone who performed a just act solely as a result of calculating that that particular action would best serve his or her own interests might be praised for being prudent, but not for having the motive constituting the merit of just acts. Furthermore, it is possible for *a sense of duty* to motivate acts of justice or fidelity that one supposes are *contrary* to one’s self-interest; but this would be impossible if the missing motive one felt to be lacking and wished to encourage were “enlightened self-interest.” Again, and even more tellingly, if someone discovered through a careful and *accurate* analysis that a particular act of secret injustice or infidelity—say, failing to return a found wallet—*would in fact* be in his or her own interest, all things considered, such a violation of the convention would then be a behavioral sign of “enlightened self-interest,” as defined; yet that motive would not bestow merit on the act. It seems, then, that although self-
But there is no inconsistency here; for although the Undoubted Maxim is derived from the First Virtuous Motive Principle, they are not equivalent. The latter requires that, for every virtuous action, there be some first virtuous motive other than duty that bestows merit on the actions. The former requires that, for every virtuous action, there be some original motive, other than duty, in human nature to motivate its performance. The motive satisfying the Undoubted Maxim need not be a virtuous motive bestowing merit on the actions; for it may instead help to create the first virtuous motive, through the creation of a convention. On the other hand, the motive satisfying the Undoubted Maxim must be, as Hume says, “original in human nature,” since it should serve to explain the origins of any relevant conventions or contrivances. The motive satisfying the First Virtuous Motive Principle, in contrast, must be virtuous in order to bestow merit on actions, but it need not be “original in human nature,” since it may arise from convention.

That is precisely what occurs in the case of justice and fidelity. Self-interest motivates the invention of the relevant conventions and the adoption of the policies of regulating one’s conduct in accordance with their rules; but the motives that we find beneficial to private and public interest, and so morally approve through sympathy, are the resulting artificial dispositions to regulate one’s conduct by the rules of justice and fidelity, dispositions that are grounded in the new convention-dependent desire to regulate one’s conduct by those rules. A particularly valuable part of these regulating dispositions is the disposition to follow the rules in a way that is independent of what might appear to be self-interested, or even publicly-interested, reasons to violate them—for the willingness to weigh costs and benefits of particular acts at the expense of the policy of uniform adherence will rightly seem dangerous to society and often to individuals as well.²⁷

### III. Can Hume Allow First Virtuous Non-Moral Motives to Justice and Fidelity?

Stephen Darwall (1993) is thus correct to propose that, for Hume, the virtuous non-moral motives constituting the virtues of justice and fidelity are the dispositions to regulate our conduct by the relevant rules. Citing passages from the Treatise concerning the motivational role of pleasure and pain, however, Darwall argues that Hume cannot consistently regard the first virtuous motives to justice and fidelity in this way, on the grounds that Hume’s conative

²⁷ Although the disposition to regulate one’s actions by the rules of justice and fidelity is the first virtuous motive bestowing merit on acts of justice and fidelity, we may well think that a fully civilized individual in modern society should also be motivated, in part, by a sense of duty in performing such actions. In the case of a natural virtue, such as affection for children, we would not, Hume holds, think less of someone who acted directly from that inclination without stopping to think of its moral character. But someone who faced a decision regarding justice or fidelity without some awareness of its moral character might well be thought to be lacking in an appropriate, and important, use of the moral sense. For because the disposition to regulate our actions by rules of justice is artificial, and requires that we discount other reasons for acting that might be attractive, it is a virtue that we should be continually attempting to re-inculcate and sustain in ourselves by practice—and this is a kind of motive of duty. Kate Abramson has helped me to appreciate the force of this point. Given this, it might be argued, one can even have a motive of meta-duty: regretting that one lacks a sufficiently strong first virtuous motive, and regretting that one doesn’t feel much present desire or ability to inculcate it in oneself by practice, either one tries to make oneself more like the kind of person who would try to inculcate it by practice, or one tries to hide one’s lack of a sense of first-order duty from oneself.
psychology cannot accommodate the regulation of conduct by rules that motivate through their own authority independent of present or prospective pleasure or pain.

Prospects of Pleasure and Pain. In support of this claim, Darwall cites three passages in which Hume comments on the motivational power of pleasure and pain:

There is implanted in the human mind a perception of pain and pleasure as the chief spring and moving principle of all its actions. (T 1.3.10.2; SBN 118)

It is 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)

Desire arises from good [i.e., pleasure] considered simply; and aversion is derived from evil [i.e., pain]. The will exerts itself, when either the good or the absence of the evil may be attained by any action of the mind or body. (T 2.3.9.7; SBN 439)

Darwall concludes from these passages that “Hume’s theory of action thus not only employs the traditional idea that the will invariably aims at the good … [but] interprets that idea hedonistically and egoistically” (1993: 422). Remarking that “just persons would appear to regard the fact that something belongs to another, or that the rule of property requires forbearance, as a motive—i.e., a ground or a reason—for not taking it” independent of the prospect of pleasure or pain in doing so, he concludes that “this motive can find no place in Hume’s theories of action and will as advanced in the Treatise” (1993: 422).

In fact, however, although Hume clearly holds that present or prospective pleasure or pain can be sufficient to move the will, he never claims that it is necessary (nor does he treat motivation by the prospect of pleasure and pain egoistically—see An Enquiry concerning the Principles of Morals Appendix 2, “Of Self-Love”). In the first passage cited, for example, he claims only that they are the “chief spring and moving principle” of human actions, not that they are the only spring or principle. The second passage cited may appear to be more restrictive, but there Hume is discussing only a particular operation of the mind: the operation by which desire or aversion for an object promising pleasure or pain gives rise to a new desire for or aversion to the inferred means to it. Hume’s point about it is only that, in such cases, it the prospect of pleasure or pain, rather than reason itself, that provides the motivational force. The third cited passage again states only that prospective pleasure or pain can be sufficient for motivation, not that such pleasure or pain is necessary. On the contrary, immediately after remarking that the will is moved by prospective pleasure or pain, he emphasizes that there are other passions that move it without any such awareness:

Beside good and evil or, in other words, pain and pleasure, 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. (T 2.3.7; SBN 439)

In any case, however, Hume need have no difficulty in explaining how human beings can be motivated to act in accordance with a rule while discounting or ignoring the prospective pleasure or pain offered by particular actions. For on his account, one can readily come to believe that

---

28 Note, for example, the omission of any mention of present pleasure and pain.
29 Darwall notices this additional claim, of course, but regards it as a further inconsistency on Hume’s part.
adopting the policy of uniform adherence is better, from the point of view of self-interest, than adopting any other course of conduct one is considering—including the default policy of simply weighing the consequences of each individual act. This belief, combined with the “calm desire” for pleasure and avoidance of pain, creates a desire to adhere uniformly—a desire that is thus caused directly by prospects of pleasure and pain. This desire may, in turn and if all goes well, motivate one to act consistently in accordance with those rules despite the potential motivating force of prospective pleasures and pains associated with particular actions. Moreover, as one realizes that the blandishments of particular pleasures and pains offered by particular violations of the rules may cause deviations from this desirable and desired policy, the desire to adhere uniformly and so not to deviate will produce a further desire to discount or ignore those blandishments. Yet it is still the consideration of the prospective pleasures and pains offered by the possible policies that motivate one to try to discount or ignore these blandishments. To be sure, these blandishments may be difficult to ignore, and they may themselves produce a compensating desire to discount or ignore the advantages of the policy of uniform adherence. The recognition of such conflicting desires about what considerations to discount or ignore may give rise to a further policy question: namely, what policy to adopt about discounting or ignoring considerations in cases covered by the rules of justice and fidelity. A consideration of prospective pleasures and pains will again tend to favor the policy of discounting individual blandishments over the policy of discounting the advantages of uniform adherence; and this may give rise to a further desire to implement the policy of discounting blandishments. Even so, uniform adherence may be psychologically difficult: it may require outside encouragement, voluntarily directed attention, habituation, repeated reinforcement (both positive and negative), and what Hume calls “strength of character” (i.e., the tendency to be motivated by “calm” rather than “violent” passions). But forces suitable for motivating uniform adherence are available in his conative psychology, and there is nothing in that psychology dictating that they cannot be effective—or, on the other hand, that they must be effective. That is as it should be, for they are not always effective.

Obligation and Authority. Darwall also suggests that Hume recognizes but cannot explain the nature of the authority that we must grant to the rules of justice and fidelity in order to regulate our actions by them independent of the pleasure and pain offered by particular actions. It is because of the tensions produced by this inability, he proposes, that Hume (1) sometimes denies that there is any non-moral motive to justice and fidelity at all, and (2) surreptitiously and somewhat confusedly introduces a third species of obligation, which Darwall calls “rule-obligation.”

We have already seen, however, that Hume does not deny that there are non-moral motives to justice and fidelity; he denies only that there are such motives naturally—i.e., without a convention. The appearance that he introduces a third species of obligation derives from a passage in which he claims:

After this convention, concerning abstinence from the possessions of others, is entered into … 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 2.1.1.22; SBN 490)

Darwall rightly emphasizes that neither natural nor moral obligation, as Hume understands them, is unintelligible without a convention; hence, he concludes, Hume here recognizes a third species of obligation that must be understood in terms of the authority of rules. If so, it should be noted,
Hume’s remark would still be odd—for the rules of justice are only some of the rules that might have such authority, and he gives no reason to think that justice must necessarily be the first convention into which one can enter. Fortunately, however, such a revisionary interpretation is not required to make sense of the quoted passage. For as Jason Baldwin (2004) has recently noted, the “obligation” that becomes intelligible with the convention of justice need not be a new species of obligation in addition to moral and natural obligation; it may instead be, and presumably is, simply a new obligation—i.e., the obligation to act justly, an obligation that is, as we have seen, both natural and moral, but is unintelligible without conventions of property.

Nor need Hume have any difficulty with explaining the authority of rules of justice and fidelity or their ability to generate “reasons” to act. “Authority,” for Hume, is simply approvable power; just as “reasons” are simply approvable considerations. The authority of the rules of justice and fidelity lies in their power to inspire the desire to act in accordance with them—a power that is regularly approved both from an interested and a moral standpoint. This approval involves a sense of natural and moral obligation—i.e., awareness that we will feel disapproval of ourselves, as self-harming and as immoral, for neglecting adherence to these rules. Since we approve of ourselves, in both an interested and a moral way, for regulating our action by these rules, they constitute reasons, both interested and moral, to act. The power and authority of these rules is particularly evident in their naturally and morally approved ability to motivate us to exclude other considerations, such as the prospects of pleasure or pain offered by particular acts, from weighing in the determination of our wills. In this way, the rules constitute what Joseph Raz has called “exclusionary reasons”—i.e., reasons to exclude other potential reasons from weighing at all in the determination of one’s action. Indeed, Raz explains the authoritative character of a directive precisely as matter of its providing a positive reason to act while excluding other reasons:

The fact that an authority requires performance of an action is reason for its performance which is not to be added to all other relevant reasons when assessing what to do, but should exclude and take the place of them” (Raz 1988: 46).

That is precisely how Hume understands the approved positive and exclusionary force bestowed by the desire to adhere to the rules of justice and fidelity when we are moved, by the consideration of prospective pleasure and pain, to adopt uniform adherence to these rules as policies.

Conclusion

I have argued that, contrary to popular perceptions, Hume does, should, and can recognize motives to justice and fidelity that satisfy the requirements of the First Virtuous Motive Principle. These motives are dispositions grounded in desires to regulate one’s actions by the rules of justice and fidelity. Because they refer to conventional rules, they are “new” motives not original in human nature; they are instead produced by self-interest, which is the original motive that satisfies Hume’s Undoubted Maxim. By creating these new motives, self-interest authoritatively restrains itself.

Another popular perception is that Humean justice and fidelity can only be maintained by erroneous judgments—what Marcia Baron (1982) has called a “noble lie.” While we cannot pursue the question further here, we have found little reason to think that any such errors are essential. Uniform adherence to justice and fidelity, both in small societies and in modern larger
ones, may be rightly judged to be the best policies to adopt from the point of view of self-interest; even if some acts of adherence fail to maximize self-interest. They may also properly win moral approval, at every stage of society, without errors about their consequences for public utility; for the moral sense may approve them even if some acts of adherence fail to maximize public interest. Justice and fidelity may indeed be judged, without error, to be the best policies—which is not, of course, any guarantee that we will not make many mistakes about them anyway.30

30 I have benefited greatly from discussions of Hume’s Circle Argument with Kate Abramson, Geoffrey Sayre-McCord, Gerald Postema, and Lorraine Besser-Jones.
References


