HAPPINESS, APPROBATION, AND RATIONAL CHOICE
STUDIES IN EMPIRICIST MORAL PHILOSOPHY

by

Hans Konrad Lottenbach

Lic.phil., University of Zurich, 1987

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

ARTS AND SCIENCES

This dissertation was presented

by

Hans Konrad Lottenbach

It was defended on

July 25, 2011

and approved by

Michael Thompson, Professor, Department of Philosophy

James Allen, Professor, Department of Philosophy

Anthony Edwards, Associate Professor, Department of Religious Studies

Dissertation Advisor: Stephen Engstrom, Professor, Department of Philosophy

HAPPINESS, APPROBATION, AND RATIONAL CHOICE

STUDIES IN EMPIRICIST MORAL PHILOSOPHY

Hans Konrad Lottenbach, PhD

University of Pittsburgh, 2011

In these studies I investigate paradigmatic empiricist accounts of three notions of moral philosophy: desire for happiness, moral approbation, and rational choice.

In the first study I situate John Locke's account of the desire for happiness in his general account of the mental faculties. I argue that in Locke's *Essay* the uneasiness of desire is to be interpreted neither as a perception of an idea nor as a volition, but as an act of a separate faculty of feeling. Only if the uneasiness of desire is understood in this way, will it be possible to make sense of Locke's claim that it constantly accompanies the perception of ideas. Understanding desire as an act of feeling will also clarify what kind of knowledge of happiness Locke assumes we have when we desire happiness.

In the second study I examine David Hume's account of the origin of the sentiment of moral approbation. Hume seems to give a general empirical explanation of this sentiment; but this explanation of the origin of moral approbation faces apparent counterexamples: the approbation of what Hume calls 'useless' or 'monkish' virtues. I argue that Hume's own treatment of these counterexamples demands a restrictive interpretation of what he labels his 'experimental method,' and an understanding of his moral philosophy as a self-enforcing genealogy of morals.

Taking as a starting point a thesis of David Gauthier's about the status of expected utility theory, I discuss – in the third study – whether an empiricist and subjectivist theory of value is

compatible with an account of rational choice that leaves room for some form of autonomy. I argue that if autonomy presupposes an activity of practical reason, the maximization of expected utility cannot be the principle of rational choice.

In each of these studies I attempt to bring into the open insufficiently acknowledged elements in empiricist moral philosophy: the role of non-experiential consciousness in Locke’s account of the universal desire for happiness, the restriction of the experimental method in Hume’s genealogy of moral approbation, and the assumption of the determinacy of the notion of expected utility maximization in Gauthier’s theory of rational choice.

TABLE OF CONTENTS

PREFACE

I thank my advisor, Steve Engstrom, and the member of my committee, Michael Thompson, James Allen, and Tony Edwards, for their support and patience. Without the friendship of Sergio Tenenbaum and Jennifer Nagel I would have been lost. This work is for Orsi.

1.0 INTRODUCTION

The central notions in empiricist moral philosophy are those of *affection* and *sentiment*. Empiricism explains both human motivation and evaluation by feelings produced by affection: all our desires, including the universal desire for happiness, as well as all our moral judgments are traced to or identified with sentiments. It is characteristic of empiricism to attempt to understand these sentiments as products of *external* affection: our active powers of movement towards happiness and moral judgment would thus be grounded in a fundamental passivity. In classical empiricism this understanding of motivation and evaluation is presented in the systematic framework of the *new way of ideas*. In contemporary versions of empiricist moral philosophy accounts of motivation and evaluation tend to be taken as parts of theories of utility and rational choice. I shall argue that paying attention to these systematic contexts reveals presuppositions in the work of empiricist philosophers – presuppositions either insufficiently made explicit or not acknowledged at all – that put into question the very nature of their moral philosophy.

In the first chapter I situate John Locke's account of the desire for happiness in his general account of the faculties of the mind. I argue that in the *Essay concerning Human Understanding* the uneasiness or pain of desire is to be interpreted neither as a perception of an idea (an act of the faculty of understanding, the faculty of perceptivity) nor as a volition (an act of the faculty of willing, the faculty of motivity), but as an act of a separate faculty of feeling (a faculty of affectivity). Only if the uneasiness or pain of desire is understood in this way, will it be possible to make sense of Locke's claim that it constantly accompanies the perception of ideas. It will become evident that, for Locke, the affection that produces desire is fundamentally not external, but a kind of *self-affection*. Understanding the central role of affectivity in his account of the mental faculties will make us recognize that in the *Essay* Locke presupposes and sometimes acknowledges a form of consciousness that is prior to all experience, i.e., prior to sensation and reflection. I argue that this non-empirical consciousness must be a kind of knowledge, and that it must include immediate knowledge of *happiness*.

In the second chapter I examine David Hume's account of evaluation or moral judgment, which is an account of the origin of a feeling: the sentiment of moral approbation. Hume seems to give a general empirical explanation of this sentiment; but this explanation of the origin of moral approbation faces apparent counterexamples: the actual approbation of virtues Hume labels 'useless' or 'monkish.' I argue that Hume's own treatment of these counterexamples (in the *Treatise*, the second *Enquiry*, and the *Essays*) demands a restrictive interpretation of what he calls his 'experimental method': the objects of the experience to which Hume appeals are only virtuous persons and their affections; that is to say, Hume presupposes that – strictly speaking – only the feelings of people with the virtuous disposition count as sentiments of moral approbation. I argue that, according to Hume, the virtuous disposition belongs only to polite gentlemen like David Hume himself. This means that the sentiment of moral approbation is quite rarely found, and that the empirical search into its origin reveals nothing universal in the human constitution. How, then, are we to interpret Hume's empiricist moral philosophy? I suggest that it is to be understood as a self-reflexive and self-enforcing genealogy of morals: it is the reflection of the virtuous man (in Hume's sense of this term, of course) on his own habit of moral sentiment, a reflection that tends to confirm this very habit.

Taking as a starting point a thesis of David Gauthier's about the status of expected utility theory, I discuss – in the third chapter – the relation between an empiricist and subjectivist theory of value and an account of rational choice that leaves room for some form of autonomy. Gauthier's work is a paradigm of contemporary attempts at presenting empiricist moral philosophy in the guise of formal theories of utility or rational choice: the pursuit of happiness is represented as the maximization of subjective expected utility. Gauthier claims that the principles of expected utility theory are the very conditions of any rational human choice and that, therefore, our fundamental practical law is a principle of happiness. He even maintains that rational choice in accordance with expected utility theory can be understood as autonomous, i.e., that – contrary to Kant – autonomous choice is choice for the sake of happiness. Gauthier presupposes that the notion of happiness interpreted as the maximization of expected utility is determinate. But I argue that, given the very nature of theories of subjective utility, the imperative of maximizing expected utility must remain indeterminate, i.e., an imperative that cannot tell us what to choose in many morally significant circumstances. Moreover, if – as suggested by Gauthier – autonomy presupposes an activity of practical reason that has as its objects desires given through our affections, the maximization of subjective expected utility cannot be the law of this activity.

2.0 LAZY LETHARGY AND FULLNESS OF JOY: LOCKE ON DESIRE AND HAPPINESS

In a remarkable passage of Chapter III of Book I of the *Essay concerning Human Understanding* Locke writes:

> Nature, I confess, has put into Man a desire of Happiness, and an aversion to Misery: These indeed are innate practical Principles, which (as practical Principles ought) do continue constantly to operate and influence all our Actions, without ceasing: These may be observ'd in all Persons and all Ages, steady and universal. (67)[1]

Only in Book II does Locke explain what this universal desire of happiness is. This explanation appears in the context of an account of the "Fountains of Knowledge" (104), an account of the origin of *ideas*. It thus often seems that in discussing *desire* Locke is discussing the origin of an idea. But this appearance is misleading. Locke can only be understood if the place of *feelings* (feelings like pleasure, pain, or desire) in his general account of the *faculties of the mind* is properly determined. I shall argue that in the *Essay* Locke distinguishes desiring both from

[1] All page references in the text are to: John Locke, An Essay Concerning Human Understanding, ed. Peter H. Nidditch (Oxford: Clarendon Press, 1975).

perceiving ideas (the operation of the faculty of the *understanding*) and from willing (the operation of the faculty of the *will*), i.e., that he is, at least implicitly, committed to the existence of a *faculty of feeling*[2] (Section 2.1). This will clarify what is actually implied in his claim that "All Men desire Happiness, that's past doubt" (279) (Section 2.2). Perhaps surprisingly, Locke will appear in the company of philosophers who thought that we can desire happiness and search after it only because we somehow know it, and that we know it only because we somehow already *have* it.[3]

2.1

Locke proposes three main theses about desire and happiness:

(1) Desire is uneasiness.

(2) We constantly desire happiness.

(3) Happiness is the utmost pleasure.

[2] For pleasure and pain I shall use the term 'feelings' rather than 'sensations'; in the *Essay*, sensation is one of the sources of *ideas*, and thus belongs to the faculty of the understanding.

[3] Paraphrasing a remark of one of these philosophers from the 17th century: we would not seek happiness unless we had already found it (Pascal, *Pensée* 553 (Brunschvicg)).

More precisely, Locke defines desire as the “uneasiness a Man finds in himself upon the absence of any thing, whose present enjoyment carries the *Idea* of Delight with it” (230).[4] That we constantly desire happiness[5] means that we remain uneasy as long as we are not happy. Since happiness is the utmost pleasure,[6] we are constantly uneasy in the absence of it. What is the utmost pleasure? Locke quotes St. Paul (1 Cor. 2, 9): “’tis what *Eye hath not seen, Ear hath not heard, nor hath it entred into the Heart of Man to conceive*” (258). Utmost pleasure is the enjoyment of God: “*With him is fullness of Joy, and Pleasure for evermore*” (258, quoting Psalm 16, 11). Locke appears to agree with no other than Augustine: *inquietum est cor nostrum donec requiescat in te.*[7] Our uneasiness comes to ease and rest only in the utmost pleasure: *fruitio dei*, the enjoyment of God.[8] But Locke offers a qualification: happiness admits of degrees. Only “in

[4] See also 251.

[5] This is stated many times in Book II of the *Essay*: 257, 259, 265, 274-5, 279, 283. (See also Locke’s essay *Of Ethic in General* (in John Locke, *Political Essays*, ed. Mark Goldie (Cambridge: Cambridge University Press, 1997), 298-99.)

[6] See 258.

[7] *Confessiones* I, I. In William Watts’s 17th century translation: “our heart cannot be quieted till it may find repose in thee” (Augustine, *Confessions I*, trans. William Watts (Cambridge, Mass.: Loeb Classical Library, 1912), 3). In his French translation of the *Essay* (a translation supervised by Locke) Pierre Coste renders “uneasiness” by the most Augustinean “inquiétude.” Malebranche’s *Recherche de la Vérité* (very well known to Locke) contains an explicitly Augustinean account of desire. Malebranche writes that in this life the soul “is always uneasy [inquiéte] because it is carried to seek what it can never find” (*Recherche* IV, II, §I in André Robinet, ed., *Oeuvres complètes de Malebranche*, tome II (Paris: J. Vrin, 1963), 17).

[8] Happiness is in the “enjoyment of him, *with whom there is fullness of joy*” (130); some other relevant passages can be found at 261, 271, 273-4, 277, 281-2.

its full extent" is it the "utmost Pleasure we are capable of," while "the lowest degree of what can be called *Happiness*, is so much ease from all Pain, and so much present Pleasure, as without which any one cannot be content" (258). But, as will become clear,[9] Locke denies that such low-degree contentment is ever without uneasiness.

According to Locke's definition, desire does not seem to presuppose some knowledge or idea of what is desired. Desire is not defined as the uneasiness upon the absence of anything *the idea of* whose present enjoyment carries the idea of delight with it. That we constantly desire happiness does not presuppose that we have an idea of happiness. By his "Historical, plain Method" (44) Locke tries to make sense of the beginning of desire. At first desire appears to be blind: we do not desire and pursue happiness because we somehow know what it is. It is because we pursue it, that we come to know it, although – in this life – not to its full extent. Our uneasiness is originally without direction: first, we are uneasy; second, the uneasiness spurs us to some action;[10] third, under favorable circumstances the action happens to hit upon what removes the uneasiness and makes us content. Through experience we may come to correlate the uneasiness, that which removes it, and the ensuing contentment. Consider a long-forgotten episode: I'm uneasy. I scream. I'm fed. I'm content. After more such episodes I will come to know what I am uneasy about, i.e., what I desire: the happiness of being fed and satiated. Another example: the city-dweller finds himself in the country and is uneasy. It happens that he returns to Paris and feels better. Sequences of events of this kind will teach him that his unease in the country is the desire to get back to the pleasures of the city.[11]

[9] See Section II of this paper.

[10] "*Uneasiness determines the Will*" (250); it is the "*spring of Action*" (252).

[11] Baudelaire returns to Paris. He feels somewhat better, but he is still uneasy, and irremediably so.

What exactly is uneasiness? How is it related to the perception of ideas? How can it operate constantly? According to the *Essay*, uneasiness is simply pain. Locke explains as follows what he means by ‘pleasure’ and ‘pain’:

> By *Pleasure* and *Pain*, I would be understood to signifie, whatsoever delights or molests us; […] Whether we call it Satisfaction, Delight, Pleasure, Happiness, *etc.* on the one side; or Uneasiness, Trouble, Pain, Torment, Anguish, Misery, *etc.* on the other, they are still but different degrees of the same thing. (128f)

Pleasure or pain is joined to (or accompanies) the perception of ideas, be it perception of ideas of sensation or of ideas of reflection:

> *Delight*, or *Uneasiness*, one or other of them join themselves to almost all our *Ideas*, both of Sensation and Reflection: And there is scarce any affection of our Senses from without, any retired thought of our Mind within, which is not able to produce in us *pleasure* or *pain*. (128)[12]

How are we to understand the accompaniment of experience (sensation and reflection) by pleasure or pain? Here are two readings of Locke’s intent:

[12] “For as in the Body, there is Sensation barely in it self, or accompanied with *Pain* or *Pleasure*; so the Thought, or Perception of the Mind is simply so, or else accompanied also with *Pleasure* or *Pain*, Delight or Trouble, call it how you please” (229). In the very first chapter of Book II Locke already points to “the satisfaction or uneasiness arising from any thought” (106). (Other relevant remarks are to be found at 110 and 537.)

(I) The *idea* of pleasure or pain is attached to almost every *idea* (of sensation or reflection).

(II) Pleasure or pain is attached to almost all *perception* of ideas (of sensation or reflection).

(I) seems closer to Locke's assertions that pleasure or pain is "join[ed] to several Thoughts" (129) or "annexed to so many other *Ideas*" (131). That Locke often writes that pleasure and pain are *joined* or *annexed* to ideas seems to fit well with his tendency to write that pleasure and pain are themselves ideas – simple ideas, to be more precise.[13] As an idea pleasure or pain would then seem to be another "Object of the Understanding when a Man thinks" (47). At one point Locke even writes that "God hath scattered up and down *several degrees of Pleasure and Pain*, *in all the things that environ and affect us*" (130). For example, to the idea of the taste of a piece of the revolting *manna*[14] would be attached an idea of pain.

It is noteworthy, however, that in Chapter VII of Book II of the *Essay* the supposed simple ideas of pleasure and pain are introduced in rather unexpected company:

> There be other simple *Ideas*, which convey themselves into the Mind, by all the ways of Sensation and Reflection, *viz.*
>
> *Pleasure*, or *Delight*, and its opposite.

[13] In Chapter VII of Book II pleasure and pain are introduced as among the "simple *Ideas*, which convey themselves into the Mind, by all the ways of Sensation and Reflection" (128).

[14] The laxative, rather than the "Manna in Heaven" that "will suit every one's Palate" (277).

> *Pain*, or *Uneasiness*.
>
> *Power*.
>
> *Existence*.
>
> *Unity*. (128)

What appears common to these notions is that they accompany all our perceptual life. In the case of the notions of existence and unity this is so because they "are suggested to the Understanding, by every Object without, and every *Idea* within" (131); in other terms, they are attached to every idea of sensation or reflection. In the case of the notion of pleasure or pain Locke insists that at any time of perceiving a degree of pain or uneasiness is attached to some idea of sensation or reflection. (Otherwise we would not be constantly in a state of desire.) The case of the notion of power is somewhat less straightforward, but no matter the intricacies of his account of power, Locke seems to maintain that to every idea is attached a notion of power: ideas received "by the impression of outward Objects on the Senses" (233) (or by the "internal Sensation" (162) of reflection) are accompanied by the notion of passive power, whereas ideas occurring "by the Determination of its [the mind's] own choice" (233) are accompanied by the notion of active power.[15]

It must be asked, however, whether certain *simple ideas* can accompany all or almost all ideas and be joined, annexed, or attached to them?[16] How is this relation of *accompaniment* to be understood? And what or who brings it about? Locke considers simple ideas the materials of all

[15] Here we can leave open the question of whether we have the notion of power from the very beginning of our perceptual life. Questions about our power over motions of our *bodies* are not our concern here.

[16] Locke appears to use these terms interchangeably.

knowledge and compares them to building blocks.[17] Now, it is easy to understand how a brick can be joined to another or several others. It is less easy to conceive a brick directly attached to most or all other bricks of a building; and it seems quite impossible to make sense of the notion of a brick joined to some or all bricks in all buildings; not to mention the utter absurdity of a brick attached to itself. But if the notions of pleasure or pain, existence, unity, and power were simple ideas and thus belonged to the materials of knowledge, they would be like the brick joined to some or all bricks in all buildings. Moreover, since the ideas of existence and unity are supposed to accompany *every* idea, they must, strictly speaking, be attached to themselves.[18] Perhaps Locke means only that each 'building' of knowledge contains, as it were, some *existence brick*, some *unity brick*, some *power brick*, and some *brick of uneasiness*. But he does not restrict his claim to *complex* ideas;[19] the accompaniment in question is supposed to apply to every idea. Locke cannot mean that whenever a 'knowledge builder' picks up, say, some simple idea of color he has to accommodate three or four other simple ideas that come with it. Moreover, it will be difficult to explain why these other simple ideas will not drag along more ideas: will not *this* idea of existence be accompanied by *this* idea of unity? And so on. (At least the many simple

[17] Simple ideas "furnish the Materials of all that various Knowledge" (132), and just as with building materials all we can do with them is "either to unite them together, or to set them by one another, or wholly separate them" (164) (that is, by *combination*, *relation*, and *abstraction*).

[18] One might, of course, reply that the idea of existence accompanies every idea *except* itself. But then, of all ideas, the idea of existence would be the only one we do not perceive as existing.

[19] It may also be pointed out that in chapter XXIII of Book II Locke ridicules the view that in some complex ideas there is, as it were, a *substance brick*.

ideas of existence and unity would also be *indistinguishable*: if someone asks what distinguishes this simple idea of unity from that, Locke cannot "send him to his Senses to inform him" (126).)

If the accompaniment is a relation of *one* (simple idea) *over many* (ideas) the supposedly simple ideas of unity, existence, power, and pleasure or pain are general or universal. In fact, Locke claims about the idea of unity that it is not only the simplest, but also the most *universal* idea:

> It has no shadow of Variety or Composition in it: every Object our Senses are employed about; every *Idea* in our Understandings; every Thought of our Minds brings this *Idea* along with it. And therefore it is the most intimate to our Thoughts, as well as it is, in its Agreement to all other things, the most universal *Idea* we have. (205)

According to Locke, all ideas are "particular in their Existence" (414) and become general or universal only by "a relation, that by the mind of Man is added to them" (414).[20] All relations are "extraneous, and superinduced" (322) and thus all universals (and – in Locke's term – all *generals*[21]) are *made*; they are "*the Inventions and Creatures of the Understanding*" (412). But, for Locke, the accompaniment of all ideas by the supposed simple ideas of unity, existence, and, perhaps, passive or active power is not the product of what he calls an *operation of the mind*. Neither is the steady accompaniment of ideas by pleasure or pain. In contrast to perceiving

[20] Locke is, of course, particularly interested in the universal or general *representation* or *signification* of ideas. He argues that this relation is added to them by an operation of the mind that includes *abstraction*. But signification is not the only possible general or universal relation of ideas.

[21] See 414.

a *complex idea* of relation, which requires the use of my active power, I am passive in perceiving the accompaniment of my ideas by the notions of unity, existence, and pleasure or pain.[22] It is, after all, not me, but the "infinite Wise Author of our being" who has been "pleased to join to several Thoughts, and several Sensations, a *perception* of *Delight*" (129). Thus, in perceiving unity, existence, power, and pleasure or pain I perceive *universals* or *generals* whose relation to all or some ideas belongs to their nature (or is instituted by God). Since universality or generality do not belong to ideas by their nature, I do not perceive an *idea* in my perception of delight or uneasiness.[23] (It is also noteworthy that in the passages about the accompaniment of experience by pleasure or pain Locke generally says neither that *ideas* of pleasure or pain are joined to ideas nor that *ideas* of pleasure or pain are joined to the perception of ideas.[24])

As a matter of fact, in the very first statement of his claim about the accompaniment of ideas by pleasure or pain, Locke does not call pleasure and pain simple ideas, but "Operations of our own Minds within" (105) where

[22] In the order of the *Essay* the account of this accompaniment comes *before* that of the operations of the mind.

[23] In the phrase "a *perception* of *Delight*" the 'of' may be taken materially rather than objectively, so that Locke is referring to a *delightful perception* rather than to a perception of the *idea* of delight. Thomas M. Lennon shows how important it is for the interpretation of Locke to give attention to the many meanings of 'of' (Thomas Lennon, "Locke and the Logic of Ideas," *History of Philosophy Quarterly* 18 (2001), 155-76.) In a *Cartesian* context the question would be this: does the idea of pleasure or pain contain *objective reality*? (Consider also the curious (and little discussed) passage in the *Sixth Meditation* (AT VII, 76) where Descartes says that the sensations of pain or pleasure are to be *distinguished* from distress of mind or delight, and that there is no intelligible, i.e., necessary, connection between them.)

[24] The exception is the already partially quoted passage at 130-31.

> the term *Operations* here, I use in a large sence, as comprehending not barely the Actions of the Mind about its *Ideas*, but some sort of Passions arising sometimes from them, such as is the satisfaction or uneasiness arising from any thought. (105-6)

Just as the operation of *perception* is not the *idea* of perception, the operation of *pleasure* is not the *idea* of pleasure. In reflection I can, of course, get the idea of perception, an idea that will itself be perceived. Similarly, I can acquire the idea of a pleasure whose perception might itself be either a pleasure or a pain. The simple ideas obtained from reflection on operations of the mind can also become the materials for the *complex* ideas that Locke calls *modes of thinking* (226) and *modes of pleasure and pain* (229).[25] (That he distinguishes these modes is a further sign that he does not take pleasure or pain to be the perception of an *idea*.) The accompaniment of perception by pleasure or pain appears thus to be the accompaniment of one operation employed about ideas (perception) by another not so employed (feeling pleasure or pain).[26] This

[25] Locke uses ‘mode’ in two senses: in the first sense (“in somewhat a different sence from its ordinary signification” (165)) it means one of the three kinds of *complex ideas* (the others being *substance* and *relation*); in the second it means a *way of being* of the mind or of an operation of the mind (e.g., perceiving rather than feeling pain). Modes in the second sense can be *observed* in reflection; modes in the first sense are not observed, but *made* by the act of the mind Locke calls *combination* (163-4). In Chapters XIX and XX of Book II of the *Essay* this ambiguity remains quite unresolved.

[26] This seems to go against Locke’s claim that “*where-ever there is Sense*, or *Perception*, *there some* Idea *is actually produced, and present in the Understanding*” (144). Thus any perception of pain would be the perception of an *idea*. But the context of this assertion is a discussion of *sensation*. Locke is only concerned to point out that *complete* sensation (sensation that does not terminate in an *impression* on the body) must include the *perception of an idea of sensation.* Whether pain is the product of *sensation* (in *Locke’s* sense of this term) is a different question.

suggests interpretation (II) of Locke’s claim in question. It will become clear that the claim can also be universalized: the operation of pleasure or pain is attached to all perception of ideas.

What, then, does Locke mean when he says that one operation accompanies (or is attached, joined, or annexed to) another? According to the *Essay*, such accompaniment occurs in many ways: *perception* accompanies *impression*;[27] *pleasure or pain* accompanies *perception*; and *volition* accompanies *pain* (*uneasiness*).[28] The relation of accompaniment is here also a relation of determination: impression determines what ideas I perceive; perception determines what I feel (pleasure or pain); and pain (uneasiness) determines what I will.

But, for Locke, there is another kind of accompaniment: the “reflex Act of Perception” (338), i.e., *consciousness*, accompanies all perception of ideas, pleasure or pain, and volition.[29] In Locke’s equivocal terms, perception of ideas is itself *perceived*: “It being impossible for any one to perceive, without perceiving, that he does perceive” (335). But the *reflex act of perception*, i.e., *consciousness* or *self-consciousness*,[30] is not itself the perception of an idea of reflection. If it were, it would have to be accompanied by another reflex act (since, by hypothesis, any perception of an idea is so accompanied). As the perception of an idea this reflex act would in turn be accompanied by yet another such act. And so on. In my reflex act of

[27] Perception “actually accompanies, and is annexed to any impression on the Body” (226).

[28] “The *will* seldom orders any action, nor is there any voluntary action performed, without some *desire* accompanying it” (256-7).

[29] Consciousness “always accompanies thinking” (335). Since the *self*, the “conscious thinking thing” (341), by consciousness “owns all the *Actions* of that thing” (341, my emphasis), it accompanies not only *perception*, but also the actions of *volition* and *feeling pleasure or pain.*

[30] Locke uses ‘consciousness’ and ‘self-consciousness’ (341) interchangeably.

perception I do not perceive the ideas of my perception, pleasure or pain, and volition; rather, I am immediately present to myself as perceiving ideas, feeling pleasure or pain, or willing things. Moreover, these operations belong to one and the same reflex act. They are *my* operations precisely because they all are united to *my* one reflex act of perception.[31]

My consciousness thus accompanies my operations in a different way from that in which my pleasure or pain accompanies my perception of ideas. The accompaniment by my reflex act of perception is *essential* to my operations.[32] In 17th century terms one could say that my reflex act accompanies my operations as their *cause of being*. Now, my operations are themselves modified: I perceive this or that idea, I feel pleasure or pain. Any co-existence, succession, or flow of perceptions and feelings is present to me only because each of the co-existing or succeeding perceptions or feelings, or any part of the flow of perception or pleasure or pain, is united to my reflex act of perception. In this way it can also become present to me that my operation of pleasure or pain accompanies my operation of perception of ideas, and that I have determinable *powers* of perception and pleasure or pain. In this accompaniment my perception of ideas is not the *cause of being* of my pleasure or pain; but, insofar as it determines what I feel, it could be called its *cause of becoming*.

Locke calls the *self* "that conscious thinking thing, […] which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern'd for it *self*" (341).[33] The very being of the self is consciousness, the reflex act of perception of its own being:

[31] Locke gives almost no hints how the reflexivity of the reflex act of perception, i.e., the reflexivity of consciousness, is to be understood.

[32] Consciousness is "inseparable from thinking, and as it seems to me essential to it" (335).

[33] One might ask how exactly Locke's notion of the *self* is related to his notion of the *mind*.

"consciousness always accompanies thinking, and 'tis that, that makes every one to be, what he calls *self*" (335). Consciousness is thus immediate perception of *existence*, *unity*, *power*, and *pleasure or pain*.[34] Furthermore, the being of the self is *being in succession*: "a train of *Ideas*, which constantly succeed one another" (182) is immediately present to my self. This consciousness of succession is immediate "for if we look immediately into our selves, […] we shall find our *Ideas* always […] passing in train, one going, and another coming, without intermission" (131).[35] *I am*, *I exist* in a constant succession of ways of relating myself to *objects*; that is, I exist in a constant succession of perceptions of ideas.[36] But I also exist in a constant succession of ways of relating myself to *myself*; that is, I exist in a constant succession of feelings of pleasure or pain. Thus, any perception of an idea is to me part of my operation of *thinking* (which includes, or can include, the perception of other ideas), and any pleasure or any pain is to me part of my operation of *feeling pleasure or pain* (which includes, or can include, other pleasures or pains).[37]

[34] It appears now that reflection on aspects of consciousness is the origin of the ideas of existence, unity, power, and pleasure or pain.

[35] We can also reflect on this succession, but "to look immediately into ourselves" and to "reflect on what is observable there" (131) are to be distinguished. Reflection, which produces an *idea* of reflection, is not an *immediate* look. Similarly, when we "find" ideas "to appear one after another" (182), this finding is not the product of *reflection*. In reflection we form the idea of succession *from* the "train of *Ideas*" (182) already found in consciousness.

[36] "Whilst we receive successively several *Ideas* in our Minds, we know that we do exist" (182).

[37] Analogously, any vital motion belonging to the "one Common Life" (331) of a brute animal is always part of an ongoing operation. A motion of inhalation, for instance, is part of the action of breathing. Moreover, the action of breathing accompanies other such actions (for example the pumping of blood).

In an important passage early in Book II of the *Essay* Locke presents the following distinction:

> The two great and principal Actions of the Mind, […], are these two:
>
> *Perception*, or *Thinking*, and
>
> *Volition*, or *Willing*.
>
> The Power of Thinking is called the *Understanding*, and the Power of Volition is called the *Will*, and these two Powers or Abilities in the Mind are denominated *Faculties*. (128)

This corresponds to his later distinction between "*Perceptivity*, or the Power of perception, or thinking" and "*Motivity*, or the Power of moving" (286).[38] There he calls perceptivity a "*Passive Power*, or Capacity" (286), such that strictly speaking its actualizations cannot be called *actions of the mind.*[39] Only motivity is an *active power*, which in minds is exercised in *actions of the mind*. Now, feeling pleasure or pain is certainly one of the *principal actions of the mind* of a self "capable of Happiness and Misery" and "concerned for it *self*" (341). Yet, as I have argued, it cannot be understood as an act of *perceptivity*, i.e., the perception of an *idea*. As will become

[38] Note that when Locke calls the power of thinking or perception the *understanding*, he uses 'understanding' in a narrow sense. In a wide sense 'understanding' (as used in the very title of the *Essay*) refers to the sum-total of mental powers ("survey'd" (46) in the *Essay*): *perceptivity*, *motivity*, and – as I shall argue – *affectivity*.

[39] Perceptivity is a passive power because it is receptivity, a "Power to receive *Ideas*" (286). In the very first chapter of Book II of the *Essay* Locke already insists that in sensation and reflection "the *Understanding* is meerly *passive*" and "cannot avoid the Perception of those *Ideas* [of sensation or reflection]" (118).

evident, Locke does not attribute it to the power of *motivity*, either. It must therefore be the actualization of another power. This faculty I shall call *affectivity* or (the power of) *feeling*.[40]

Affectivity or the faculty of feeling stands – as it were – in between the faculties of understanding and will: we determine our *will* to motion or thought by the act of our *feeling*, i.e., the uneasiness that accompanies the act of the *understanding*, i.e., the perception of ideas. Locke insists on distinguishing will and desire:

> Whence it is evident, that *desiring* and *willing* are two distinct Acts of the mind; and consequently that the *Will*, which is but the power of *Volition*, is much more distinct from *Desire*. (250)

For Locke, "the reason why the *will* and *desire* are so often confounded" (257) is simply the fact that volition is always accompanied by the uneasiness of desire. But what is true of willing and desiring also applies to desiring and thinking: they are distinct acts of the mind and belong to different faculties. In a curious passage Locke imagines what our state would be if we lacked feelings of pleasure or pain:

[40] There is a faculty/act ambiguity in 'feeling' (just as in the French 'sentiment' and the German 'Gefühl'). I shall use 'feeling' both for the faculty (as the term is used in the title of a once famous novel: *The Man of Feeling*) and for its act. The context should make things clear. (An early hint of a tri-partite division of faculties in the *Essay* is a passage near the beginning of Book II where Locke ascribes to the soul "Thinking, Enjoyments, and Concerns" (110).

> And so we should neither stir our Bodies, nor employ our Minds; but let our Thoughts (if I may so call it) run a drift, without any direction or design; and suffer the *Ideas* of our Minds, like unregarded shadows, to make their appearances there, as it happen'd, without attending to them. In which state Man, however furnished with the faculties of Understanding and Will, would be a very idle unactive Creature, and pass his time only in a lazy lethargick Dream. (129)[41]

In this state the will would be an unused *bare faculty*[42] since the understanding alone would not be sufficient to determine it. This implies that without a faculty of feeling distinct from both the will *and* the understanding we could not be the active creatures we are. (In general, Locke claims that "the *Idea* in the mind of whatever good, is there only like other *Ideas*, the object of bare unactive speculation; but operates not on the will, nor sets us on work" (255). If, like Berkeley,[43] he thinks that *all* ideas are only the objects of such 'unactive speculation,' he cannot hold that pleasure and pain are ideas. At any rate, when he contrasts the 'unactive'

[41] The passage recalls Hobbes's account of one sort of "Trayne of Thoughts": "The first is *Unguided, without Designe*, and inconstant; Wherein there is no Passionate Thought, to govern and direct those that follow, to it self, as the end and scope of some desire, or other passion" (*Leviathan* I.III; in Thomas Hobbes, *Leviathan*, ed. C.B. Macpherson (Harmondsworth: Penguin, 1968), 95).

[42] According to Leibniz, talk about a *will* in such a state would be unintelligible. For Leibniz's complaints about Locke using the incomprehensible notion of a *bare faculty* or *bare power* see his *New Essays on Human Understanding* 2.1.2 and 4.3.6; in G.W. Leibniz, *New Essays on Human Understanding*, trans. and ed. Peter Remnant and Jonathan Bennett (Cambridge: Cambridge University Press, 1982), 110 and 379.

[43] See *A Treatise Concerning the Principles of Human Knowledge* 25, in *The Works of George Berkeley*, vol. 2, ed. A.A. Luce and T.E. Jessop (London: Thomas Nelson, 1949), 51-52.

perception of the idea of some good with the uneasiness that ‘sets us on work,’ he never describes the uneasiness as the perception of an idea.)

Is affectivity an active or a passive power? Insofar as we are “*sensible* […] of Pleasure and Pain” (341), insofar as we receive the modifications of pleasure or pain, we are as passive as in perception.[44] But in pleasure or pain we do not receive ideas. Pleasure and pain are *feelings*; they belong to “some sort of Passions arising sometimes from them [ideas or – more precisely – perceptions of ideas]” (106).[45] As a matter of fact, “satisfaction or uneasiness” is such a passion that may be “arising from any thought” (106). We can say that Locke distinguishes two forms of sensibility: *perceptivity*, actualized in the perception of ideas of sensation and reflection from *experience*; and *affectivity*, actualized in feelings from *self-affection*.[46] *Reflection* (or “internal Sensation” (162)) might also be called *self-affection*; but it is essential to distinguish between the “Fountains” (104) of the perception of ideas and the fountain of feeling. (It seems that in the self-affection of *reflection* one *part* of the self (the self perceiving an idea of sensation) affects another (the self perceiving an idea of perceiving an idea of sensation), whereas in the self-

[44] “For in bare naked *Perception*, the Mind is, for the most part, only passive; and what it perceives, it cannot avoid perceiving” (143).

[45] These passions belong to the “Operations of our own Minds within” (105), and – in a broad sense – can even be called *actions.*

[46] When *Locke* uses the term ‘sensibility,’ he usually means the susceptibility to pleasure or pain: “he who made us […] will restore us to the like state of Sensibility in another World, and make us capable there to receive the Retribution he has designed to Men” (542). What makes us capable of receiving retribution is the capacity of feeling pleasure or pain.

affection of *feeling* the *whole* self affects the *whole* self.[47]) Insofar as affectivity is a faculty of *self-affection*, it might also be considered an active power, although it should not be confused with the active power of motivity: The motivity of the mind (its *will*) is exercised in moving bodies and ‘moving’ thoughts (for instance, in bringing “into view *Ideas* out of sight, at one’s own choice, and to compare which of them one thinks fit” (286) and similar operations requiring voluntary attention), whereas the affectivity of the mind (the *self*) is actualized in the self affecting itself.[48]

2.2

What are the consequences of all this for Locke’s account of desire and happiness? That desire is uneasiness, i.e., pain, implies that in desire the self is conscious of its operation of *feeling*. At least *in this life*, desire is constant because no moment of the operation of feeling is ever without uneasiness. Locke writes: “we finding imperfection, dissatisfaction, and want of complete happiness, in all the Enjoyments which the Creatures can afford us, might be led to seek it in the enjoyment of him, *with whom there is fullness of joy, and at whose right hand are pleasures for*

[47] Again, Locke gives no further explanation of this reflexivity. (Is the whole affecting itself a *simple* whole? Is it like an organic whole?)

[48] Although Locke for the most part seems to assume that “Pleasure or Pain follows upon the application of certain Objects to us, *whose Existence we perceive*” (537, my emphasis), it is so far an open question whether self-affection *necessarily* requires the perception of ideas.

evermore" (130).[49] Now, in the uneasiness of desire I feel myself as someone who could be pleased (either to the full extent or to a lesser degree): in consciousness I am "*sensible*, or conscious of Pleasure and Pain, *capable* of Happiness or Misery" (341, my emphases). In the uneasiness of desire I also feel myself *concerned* for myself: "Happiness and Misery, being that, for which every one is concerned for *himself*" (341f). In concern for myself I am not only "conscious of Pleasure and Pain, capable of Happiness and Misery," but I also feel misery as my imperfect, happiness as my perfect being – if indeed I ever come to feel happiness. That nobody is "feeling pain, that he wishes not to be eased of" (251) requires that everybody feel he *could* and *should* be happy: "A concern for Happiness" is "the unavoidable concomitant of consciousness, that which is conscious of Pleasure and Pain, desiring, that that *self*, that is conscious, should be happy" (346).

Does this consciousness of uneasiness as *imperfection* or *want* presuppose some consciousness of happiness? Some consciousness of *actual* happiness? For how could I feel my imperfect being unless there were in me some consciousness of my more perfect being which enables me to recognize my defect by comparison?[50] This consciousness of my more perfect

[49] In this life even in *joy* we remain uneasy: "the present moment not being our eternity, whatever our enjoyment be, we look beyond the present, and desire goes with our foresight" (257). Hobbes, too, claims that "there is no such a thing as perpetuall Tranquillity of mind, while we live here; because Life it selfe is but Motion, and can never be without Desire" (*Leviathan* I, VI; ed. Macpherson, 129-30). But for Hobbes this means that a notion like that of the *utmost pleasure* is "as incomprehensible" to us "as the word of School-Men *Beatificall Vision* is unintelligible" (*Leviathan* I, VI; ed. Macpherson, 130).

[50] I am, of course, paraphrasing a passage from Descartes's *Third Meditation* (AT 45-6). (Locke was probably also familiar with this text: "Now, that we have an *Idea* or *Conception* of *Perfection*, or a *Perfect Being*; is Evident, from the *Notion* that we have, of *Imperfection* so familiar to us: *Perfection* being the *Rule* and *Measure* of *Imperfection*,

being, the *standard* or *measure* of my *feeling*, cannot be the perception of an idea of happiness (in Locke's sense of 'idea'). Locke obviously assumes that we have some idea of happiness, a complex idea we form by the mental operations of *enlarging* and *abstraction* applied to simple ideas of pleasures received in reflection.[51] But the measure of feeling is *happiness*, rather than some *idea* of happiness. I cannot recognize that my uneasiness is an imperfection by comparing it with an idea of happiness, especially an abstract idea.[52] In order to feel uneasiness it cannot be necessary that I have already reflected on my feelings and started forming more or less elaborate abstract ideas or thoughts about happiness.[53]

Moreover, the measure of my feeling cannot be an idea derived from what it is supposed to measure: In first forming the idea of happiness I would have to start with reflection on feelings

and not *Imperfection* of *Perfection*" (Ralph Cudworth, *The True Intellectual System of the Universe* (London: Royston, 1678), 648).)

[51] According to Locke the operation of *enlarging* as applied to ideas is a kind of *composition*: "Under this [operation] of Composition, may be reckon'd also that of *ENLARGING*; wherein though the Composition does not so much appear as in more complex ones, yet it is nevertheless a putting several *Ideas* together, though of the same kind" (158). To talk about an *enlarged idea* is thus to talk about a *complex idea* (or a *complex of ideas*).

[52] The perception of the idea of happiness may itself be accompanied by pleasure. But obviously *this* pleasure cannot be the measure of feeling. Locke reports that in reflection he found "that the expectation of eternal and incomprehensible happiness in another world is that also which carries a constant pleasure with it" ("Thus I Think," in *Political Essays*, ed. Mark Goldie (Cambridge: Cambridge University Press, 1997), 297).

[53] Jean-Jacques Rousseau will accuse philosophers of assuming that in order to live I need to be "very great reasoner and a profound metaphysician" (*Discourse on the Origin of Inequality*, Preface; in Jean-Jacques Rousseau, *Oeuvres complètes III*, ed. Marcel Raymond et Bernard Gagnebin (Paris: Gallimard, 1964), 125). Scandalously, he claims that for us "the state of reflection is a state against nature" (Rousseau, *Oeuvres complètes III*, 138). But *reflection* (as understood by Locke) certainly cannot be our *beginning* in life (in perceiving, feeling, and willing).

which I do not yet recognize as feelings of imperfection or perfection, uneasiness or satisfaction (since – by hypothesis – I do not yet have their measure). Whatever idea I would form in this way would be neither the idea of happiness nor that of misery.[54] This is not to deny that there could be reflective ideas of happiness and misery, but they would have to be derived from feelings of pleasure and pain already recognized as feelings of perfection and imperfection.

It is also worth pointing out that in forming ideas of pleasure or happiness *via* reflection I can go astray. The products of reflection ("properly enough […] call'd internal Sense" (105)) are simple ideas of reflection whose *agreement with the reality of things* is as questionable as that of simple ideas of sensation. Even if simple ideas of reflection, like simple ideas of sensation, "*are not fictions* of our Fancies" (564), because they are produced by causes in "the reality of Things" (563), nothing is thereby established about whether they resemble them in any way. Strange as it may sound, the idea of my uneasiness might misrepresent my real feeling. Similarly, my complex idea of my happiness, which – in Locke's oddly Malebranchean terms – is not its own *archetype*,[55] might misrepresent my state of feeling. Nothing in Locke's *way of ideas* rules out that my complex *reflective* ideas can be like the "Reveries of a crazy Brain" (563).[56]

[54] The *enlarged* and *abstract* idea obtained in this way might be called the idea of *more feeling*. Perhaps we are here at the origin of some form of *utilitarianism.*

[55] A complex idea that is its own archetype is "not designed to represent any thing but it self" and therefore never "capable of a wrong representation" (564).

[56] This is almost never noticed in Locke commentary. (For an exception see Martha Brandt Bolton, "The Taxonomy of Ideas in Locke's *Essay*," in *The Cambridge Companion to Locke's "Essay concerning Human Understanding*, ed. Lex Newman (Cambridge: Cambridge University Press, 2007), 85-6.) (It is remarkable that in his chapter on the *reality of knowledge* Locke considers neither simple ideas of *reflection* nor complex ideas made of them that may be "supposed Copies" (568) of operations of the mind.) Moreover, according to Locke, these reflective *reveries* or

For Locke, what then could be the consciousness that enables me to recognize my uneasiness as a defect in my being if it is not the perception of a reflective idea? On page 618 (!) of the *Essay* Locke writes that "nothing can be more evident to us, than our own Existence," and that "*we have an intuitive Knowledge of our own Existence*, and an internal infallible Perception that we are."[57] This is neither the perception of an idea of sensation nor of reflection. We know our existence not by having an idea of it, but by "that consciousness, which is inseparable from thinking" (335).[58] But, according to Locke, it is not essential to consciousness that it be modified by perceptions of ideas. That consciousness is inseparable from thinking does not imply that

"Fancies" (563) will not be harmless: someone who is mistaken in his thoughts about his happiness has gone astray in "his own Thought and Judgment, what is best for him to do" (264). More dramatically: "He has vitiated his own Palate, and must be answerable to himself for the sickness and death that follows from it" (271).

[57] Locke hints at this intuitive knowledge in an important passage from Book II: "Every act of sensation, when duly considered, gives us an equal view of both parts of nature, the Corporeal and Spiritual. For whilst I know, by seeing or hearing, *etc.* that there is some Corporeal Being without me, the Object of that sensation, I do more certainly know, that there is some Spiritual Being within me, that sees and hears" (306).

[58] On this point Locke agrees with Malebranche: "We do not know it [the soul] at all by its idea: […] we know it by *consciousness* [*conscience*]" (*Recherche de la Vérité* III, II, VII, §IV; in *Oeuvres complètes de Malebranche,* tome I, 451). (That for Locke the *intuition* of our own existence just *is* consciousness is further supported by some of his remarks on Descartes's proof of the existence of God: "our own existence is known to us by certainty yet higher than our senses can give us of the existence of other things, and that is internal perception, a self-consciousness, or intuition" (Lord Peter King, *The Life of John Locke* (London: Colburn and Bentley, 1830), vol. II, 138-9.) It is beyond the scope of this essay to address the question of how intuitive knowledge of my existence can be "*the Perception of the Agreement or Disagreement of two* Ideas" (525). Under any interpretation, it will be difficult to understand the notions of myself and my existence as (simple) ideas derived from experience.

thinking is inseparable from consciousness.[59] (Notoriously, Locke rejects the "Opinion, that the Soul always thinks" (108); he denies "that actual thinking is as inseparable from the soul, as actual Extension is from the Body" (108).[60]) It is also not essential that our *affectivity* be determined by the perception of ideas. After all, Locke thinks that we are "intended for a State of Happiness" (277) which is nothing but "the enjoyment of him, *with whom there is fullness of joy*" (130, quoting Ps. 16: 11). However consciousness is to be characterized in this state (and Locke is too modest to speculate much about this), it is not a thinking consciousness, and its act of affectivity, the fullness of joy, is not determined by the perception of ideas of sensation or reflection.

The state of *fullness of joy* can be compared to the state (already briefly considered above) in which "Man, however furnished with the faculties of Understanding and Will, would be a very idle unactive Creature, and pass his time only in a lazy lethargick Dream" (129), a state I shall simply call *lazy lethargy*. In both states faculties we find *in this life* are supposed to be absent (or only bare faculties). It might seem that in *lazy lethargy* there is neither affectivity nor motivity, whereas in *fullness of joy* there is only affectivity. But this needs to be qualified. In *fullness of joy* there might be a form of perceptivity distinct from that actualized in the perception

[59] To prevent misunderstandings: 'thinking' is here used in *Locke's* sense (as referring to the perception of ideas or sensation or reflection). In a *Cartesian* context the claim would appear to be nonsense.

[60] Locke very charmingly confesses that he has "one of those dull Souls, that doth not perceive it self always to contemplate *Ideas*" (108). In these passages we might substitute 'consciousness' for 'soul.' Locke had better not say that the soul may (sometimes) be only a bare faculty. As Leibniz points out correctly, this would make the soul belong to "mere fictions, unknown to nature, and obtainable only by abstraction" (G. W.Leibniz, *New Essays on Human Understanding*, 2.1.2; trans. Peter Remnant and Jonathan Bennett, 110). Less politely, the soul would (sometimes) be nothing (and Locke in his dullness would be dead).

of *ideas* of sensation or reflection (some *immediate vision of God*) and a form of will distinct from the will determined by the uneasiness accompanying perceptions of *ideas* of sensation or reflection (some *will willing what it already has*[61]). Similarly, nothing in Locke's description of *lazy lethargy* rules out that in this state there might be a feeling or a will *independent of any perception of ideas.*

How could one make sense of feeling or willing in *lazy lethargy*? Lazy lethargy could be my state, my existence, or my life only if my preservation depended in no way on my being conscious of any pleasure or pain arising from my thoughts. (In this respect it would, of course, be similar to the *fullness of joy*.) But, again, my feeling could be determined by something different from my thoughts. What could that be? In lazy lethargy there appears to be a possible source of feeling: my very existence, my life. I am conscious of it and feel no lack or imperfection in it. About lazy lethargy one can't help asking: wouldn't it be sweet? If indeed it is *pleasant* or *happy*, it is so through an *immediate* self-affection without any detour *via* the perception of ideas. Therefore, such pleasure would be in no way affected by their changes and be as permanent as my existence. It might even be said to determine my will. In lazy lethargy I would be a "very idle unactive Creature" (129) only because I would not at all be concerned for things appearing in the perception of ideas and thus would do nothing about them. Still, I would be concerned for myself, for the existence and life of which I am conscious. That I already *have*

[61] For a widely read 17th century account of this old notion see François de Sales' *Traité de l'Amour de Dieu* V, III (François de Sales, *Oeuvres*, ed. André Ravier (Paris: Gallimard, 1969), 572-76).

it does not rule out that I *will* it. My will which would will nothing that appears in perception, i.e., nothing different from my existence, would then be a will that wills *itself*.[62]

Is this more than idle speculation, a lazy dream? To return to our main topic, Locke needs to explain how the uneasiness of desire can be the consciousness of imperfection: by what consciousness of happiness do I measure my feeling? Since it cannot be the perception of an idea, it must be a more immediate consciousness. What could it be? Locke's hints about *fullness of joy* and *lazy lethargy* suggest an answer: the measure of my feeling is the happy consciousness of existence.

Locke begins Book II of the *Essay*, i.e., his "true *History of the first beginnings of Humane Knowledge*" (162), with the phrase "Every Man *being conscious to himself,* That he thinks" (104, my emphasis). Indeed, if originally I were not conscious to myself (in a reflex act of perception of my being), I could not come to perceive ideas; and if originally I were not feeling happiness, I could not come to feel uneasiness on the occasion of the perception of ideas. Speaking historically, one might say that, according to Locke, I could not have become uneasy unless I had been happy before, or, more lyrically, unless I had been *born happy*.[63]

[62] Locke seems to recognize that there could be a will when there is no uneasiness at all: "When a Man is perfectly content with the State he is in, which is when he is perfectly without any *uneasiness*, what industry, what action, what *Will* is there left, but to continue in it?" (252). He presents this, however, as a fact of which "every Man's observation will satisfy him" (252), and thereby contradicts his claim that *in this life* we never find ourselves perfectly without uneasiness.

[63] Here it is perhaps not superfluous to point out that when Locke asks the question "*at what time a Man has first any* Ideas" (108) he is *not* asking the question: "at what time does a man first become conscious?"

In structural rather than historical terms, uneasiness, as the feeling of imperfection, is a feeling of a difference: the difference between my imperfect and perfect being. How can I feel this difference? Locke claims we are conscious of difference by an operation of the mind he calls *discerning*: it is by discerning that “the Mind […] perceives two *Ideas* to be the same, or different” (156). He points out that “it is the first act of the Mind, (without which, it can never be capable of any Knowledge,) to know every one of its *Ideas* by it self, and distinguish it from others” (592). Since earlier in the *Essay* (in the main discussion of *operations of the mind*) *perceiving* is called the *first act of the mind* “about our *Ideas*” (143), discerning must – strictly speaking – be a *way* of perceiving ideas, i.e., distinct perceiving.[64] In distinct perception I am immediately conscious that the idea I perceive is *not* another one I perceive or did perceive (and could perceive again).[65] This consciousness is immediate since I do not need to form an idea of difference in order to perceive distinctly. Now, some form of discerning must apply not only to the perception of ideas, but also to feeling. In uneasiness it is immediately present to me that in my operation of feeling my current degree of pleasure is *not* the utmost pleasure I actually felt (and could feel again), the pleasure in my existence lacking nothing.

[64] Locke ascribes the task of *noticing* bodily *impressions* by means of the perception of an idea both to the “discerning Faculty” (132) and to the faculty of perception (143). (Since his full discussion of *noticing* is in the chapter on the faculty of *perception*, the earlier passage may be interpreted as emphasizing an important *aspect* or *mode* of perception.)

[65] A perceiver “can never be in doubt when any *Idea* is in his Mind, that it is there, and is that *Idea* it is; and that two distinct *Ideas*, when they are in his Mind, are there, and are not one and the same *Idea*” (592). For ideas, one way of *being in the mind*, is, of course, *succession*.

According to Locke, that *in this life* we constantly desire happiness is a consequence of the fact that our relation to the *objects* of *this world*, i.e., the perception of ideas, constantly produces some uneasiness (a consciousness of reduced pleasure). In other terms: the desire for happiness is a consequence of our affectivity being sensibly affected (where this may include being affected by the internal sense of reflection). In response to the question of why this should be so, Locke tells us to admire "the Wisdom and Goodness of our Maker, who designing the preservation of our Being, has annexed Pain to the application of many things to our Bodies, to warn us of the harm that they will do" (129f). But the preservation of our sensible existence is not all the design of our Maker. "The want of complete happiness, in all the Enjoyments which the Creatures can afford us" (130), i.e., the uneasiness that in varying degrees accompanies all our perception of *ideas*, is an indication that our happiness does not lie in *thought* or *knowledge*. In a notorious passage Locke writes:

> For since the Things, the Mind contemplates, are none of them, *besides it self* [my emphasis], present to the Understanding, 'tis necessary that something else, as a Sign or Representation of the thing it considers, should be present to it: And these are *Ideas*. (720f)

To say that I am not satisfied in *thought* is thus to say that I am not satisfied with the presence of *ideas*. I can be satisfied only by the presence of *things.*[66] In the *fullness of joy* I would be

[66] We do not need to take a position on the endlessly debated question of what exactly Locke means by 'representation' and 'presence to the mind.' It suffices here to point out that any answer must take into account that for Locke there are two distinct forms of presence to the mind: one somehow involving *ideas* or *appearances* (287),

satisfied by the presence of God. But even an inkling of this possibility presupposes an actual satisfaction in the presence of another ‘thing’: my own existence.[67]

the other not. (That we are satisfied only in the presence of *things* means that we are satisfied only by *truth* – even if putting it this way might add too much *unction* to the claim. *Locke* for his part says that truth is “that which all Mankind either do, or pretend to search after” (574). He proceeds, however, to assert that “*Truth* then seems to me, in the proper import of the Word, to signify nothing but *the joining or separating of Signs, as the Things signified by them, do agree or disagree one with another* (574). It may be doubted that all mankind searches, or pretends to search, after *that*.)

[67] This actual pleasure in one’s own existence could be called – in a term that would horrify Augustine – *fruitio sui*. As a matter of fact, in a great work known to Locke it is so called: “it is certain that without *Consciousness* […] nothing can be Happy (since it could not have any *Fruition* of it self)” (Ralph Cudworth, *The True Intellectual System of the Universe* (London: Royston, 1678), 847. (On Cudworth and British accounts of consciousness in the 17th century see Udo Thiel, *Lockes Theorie der personalen Identität* (Bonn: Bouvier, 1983), 67-104.)

3.0 MONKISH VIRTUES, ARTIFICIAL LIVES: ON HUME'S GENEALOGY OF MORALS

> The merchant's toil, the sage's indolence,
> The monk's humility, the hero's pride,
> All, all alike, find Reason on their side.
>
> (Alexander Pope, *An Essay on Man*, Epistle II, 172-4)

Hume's moral philosophy is often interpreted as an example of naturalistic approach to ethics. J.L. Mackie, for instance, writes that in Hume the questions of moral philosophy are answered "in sociological and psychological terms, by constructing and defending a causal hypothesis."[68] Similarly, Páll S. Árdal claims that Hume "is concerned with an attempt to discover those psychological laws that explain human emotions (including moral emotions) and the behavior of people in society."[69] I argue in this essay that if Hume is read in this way as developing a general explanatory theory of moral sentiments, he faces an inescapable dilemma. Section 3.1 presents

[68] J.L. Mackie, *Hume's Moral Theory* (London: Routledge & Kegan Paul, 1980), 6.

[69] Páll S. Árdal, *Passion and Value in Hume's Treatise* (Edinburgh: Edinburgh University Press, 1989), 2. In the same spirit, Barry Stroud holds that in all his philosophy Hume is concerned with a "completely comprehensive empirical investigation and explanation of why human beings are the way they are, and why they think, feel, and behave as they do" (Barry Stroud, *Hume* (London: Routledge & Kegan Paul, 1977), 224).

the dilemma. In Sections 3.2 and 3.3, I argue why for Hume – interpreted as a proponent of general psychological laws – there is no way out of this dilemma. In Sections 3.4 to 3.6, I discuss an alternative reading of Hume according to which he is concerned with psychological laws only insofar as they reflect a uniquely *natural* standard of virtue and vice. The problems that still emerge if Hume is interpreted in this way will provoke some conjectures (developed in Sections 3.7 and 3.8) about the nature of Hume's project in moral philosophy. I shall sketch an interpretation of Hume which allows us to understand his moral philosophy not as a general explanatory theory, but as a particular kind of *genealogy* of morals. Throughout this essay I am concerned with the often noticed problem that – as Árdal puts it – "there is no neat division between Hume's psychology and his moral theory."[70] The interpretation sketched here can explain why the absence of such a neat division in Hume's writings does not indicate a fundamental confusion.

3.1

In the *Enquiry Concerning the Principles of Morals*,[71] Hume summarizes his investigations in moral philosophy as the endeavor

[70] Árdal, 2.

[71] From now on, simply *Enquiry*. References to Hume's works are given according to the following abbreviations:

T: *A Treatise of Human Nature*, 2nd ed. L.A. Selby-Bigge and P.H. Nidditch, eds. (Oxford: Clarendon Press, 1978).

E: *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd ed. L.A. Selby-Bigge and P.H. Nidditch, eds. (Oxford: Clarendon Press, 1975).

AD: *A Dialogue* (appended to the *Enquiries*; see edition referred to above).

> to collect, on the one hand, a list of those mental qualities which are the object of love or esteem, and form a part of personal merit; and on the other hand, a catalogue of those mental qualities which are the object of censure and reproach, and which detract from the character of the person possessed of them; subjoining some reflections concerning the origin of these sentiments of praise or blame. (*E*, 312)

Hume argues that the source of moral approbation is the sentiment of humanity, i.e., the sentiment for whatever is useful or agreeable either to ourselves or others. A mental quality is a *virtue* because we approve of it on the basis of its utility or agreeableness. With his list of morally praiseworthy qualities Hume intends to represent virtue without a “dismal dress” (*E*, 279), so that

> nothing appears but gentleness, humanity, beneficence, affability; nay, even at proper intervals, play, frolic, and gaiety. She [virtue] talks not of useless austerities and rigours, suffering and self-denial. She declares that her sole purpose is to make her votaries and all mankind, during every instance of their existence, if possible, cheerful and happy; nor does she ever willingly part with any pleasure but in hopes of ample compensation in some other period of their lives. The sole trouble which she demands, is that of just calculation, and a steady preference of the greater happiness. And if any austere pretenders approach her, enemies to joy and pleasure, she either rejects them as

ES: *Essays Moral, Political, and Literary*, Eugene F. Miller, ed. (Indianapolis: Liberty Classics, 1985).

D: *Dialogues Concerning Natural Religion*, Norman Kemp Smith, ed. (Indianapolis, Bobbs-Merrill, 1947).

> hypocrites and deceivers; or, if she admit them in her train, they are ranked, however, among the least favoured of her votaries. (*E*, 279-80)

However, this pretty picture notwithstanding, it seems a fact that some people do not approve of useful or agreeable mental qualities *because* they see them as useful or agreeable.[72] In his *Essays* Hume acknowledges perspectives on (moral) approbation that "form themselves naturally in the world" (*ES*, 138n.), but do not seem to be solely the result of a recognition of useful and agreeable mental qualities.[73] As Hume admits, advocates of religious morality – 'austere pretenders' as he calls them – even show approbation of character traits that appear manifestly useless and disagreeable: "Celibacy, fasting, penance, mortification, self-denial, humility, silence, solitude, and the whole train of monkish virtues" (*E*, 270). All this seems to contradict Hume's frequent assertions that about moral questions there is almost complete unanimity among mankind.[74]

[72] Some people approve of justice even if the *circumstances of justice*, i.e., the circumstances (discussed by Hume in *T*, 484-501) under which justice is useful, are absent (and they may do so not because of a habit first acquired in these circumstances). For instance, some people, will be shocked by the "gentle usage" Hume envisions for "a species of creatures intermingled with men, which though rational, were possessed of such inferior strength, both of body and mind, that they were incapable of all resistance, and could never, upon the highest provocation, make us feel the effects of their resentment" (*E*, 190).

[73] See the sequence of essays that contains *The Epicurean*, *The Stoic*, and *The Platonist* (*ES*, 138-158). As will become clear, in these essays Hume parodies perspectives on moral approbation that differ from his own.

[74] E.g., *E*, 174-75; *T*, 552; *ES*, 166n.3.

Given these phenomena, Hume seems to face a dilemma. Either the approbation for 'useless,' 'disagreeable,' or even 'monkish' virtues is not really a *moral* approbation, or the claim that moral approbation arises from useful and agreeable qualities is false.

Thus Hume needs either an explanations of how a sentiment of approbation can seem to be a sentiment of moral approbation when, in fact, it is not, or an account of the origin of moral approbation that is richer than the one sketched in the *Enquiry*. We shall call this problem the 'dilemma in Hume's moral sentimentalism.'

Hume seems to acknowledge the dilemma in *A Dialogue*, which forms a kind of appendix to the *Enquiry*. There he appears to grant that, for example, the approbation given by many people to the character of Blaise Pascal is moral approbation although Pascal exhibits the monkish virtues in the most extreme fashion. Hume claims that this approbation is due to the illusions of "religious superstition or philosophical enthusiasm," rather than to "the natural principles of the mind."[75] He adds that "when men depart from the maxims of common reason, and affect these artificial lives, no one can answer for what will please or displease them" (*AD*, 343).

Hume's use of 'natural' and 'artificial' in this context is revealing. Hume is generally very careful in distinguishing different meanings of 'natural.'[76] A quality, institution, or sentiment is natural in the first sense ('natural' as opposed to 'miraculous') if its occurrence is explainable by general principles of human nature; it is natural in a second sense ('natural' as

[75] *AD*, 343; see also *E*, 270: "And as every quality which is useful or agreeable to ourselves or others is, in common life, allowed to be a part of personal merit; so no other will ever be received, where men judge of things by their natural, unprejudiced reason, without the delusive glosses of superstition and false religion."

[76] *T*, 473-75.

opposed to ‘unusual’) if it is frequently encountered. In both theses senses of ‘natural,’ however, moral approbation of Pascal’s character is perfectly natural. It often occurs and can be explained by reference to the regular effects of certain types of moral education or particular experiences in life. In a third meaning of ‘natural’ Hume opposes ‘natural’ to ‘artificial’: a sentiment is natural in this sense if it arises immediately, as it were instinctively, without the intermediary artifices of “reason and forethought.”[77] But it seems Hume can hardly object to the approbation of certain qualities because it is the result of an ‘artificial life,’ i.e., a life in which sentiments of praise and blame are mediated by reason and forethought. After all, according to his own account of moral sentiments, the approbation of the virtue of justice is not natural, but artificial.[78] Thus, in calling the esteem for monkish virtues unnatural Hume seems to presuppose a notion of *appropriate* approbation. It seems unclear, however, whether such a notion is available to him.

Hume is interested in the “true origin of morals” and believes it can be found only by following a “very simple method” (*E*, 173):

> We shall analyze that complication of mental qualities, which form what, in common life, we call Personal Merit: we shall consider every attribute of the mind, which renders a man an object either of esteem and affection, or of hatred and contempt; every habit or sentiment or faculty, which, if ascribed to any person, implies either praise or blame, and may enter into panegyric or satire of his character and manners. (*E*, 173-74)

[77] *E*, 307-8n.2.

[78] As we shall see, even the approbation of the virtues Hume calls ‘natural’ involves some artifices.

Given this “experimental method” (*E*, 174), how can approbation of monkish and other virtues not listed in Hume’s catalogue[79] be discredited *experimentally* as “delusive glosses of superstition” (*E*, 174)? How can Hume exclude from the catalogue of the virtues any mental qualities – as monkish as they might be – that are in fact praised and do enter into panegyrics? Hume’s concern seems to be not only the true origin of morals, but also the origin of true morals: “The end of all moral speculations is to teach us our duty” (*E*, 172).

We might add that it is remarkable how often Hume states claims that seem to be general in scope only to present strong counterexamples to these very claims, without seeming in the least bit concerned that these counterexamples amount to refutations. The most famous case is, of course, Hume’s theory of the origin of ideas and the counterexample – acknowledged by Hume as indeed a counterexample – of the missing shade of blue.[80] Another clear example comes from his social and economic thought: in the essay *Of Refinement in the Arts*,[81] Hume states his position in the great 18th century debate about the relation between luxury and virtue: he takes it to be a general truth he can prove “that the ages of refinement are both the happiest and most virtuous” (*ES*, 269); but this does not prevent him from emphasizing the extraordinary virtue in the early (and very frugal) Roman Republic. Similarly, in our case of the origin or morals: to the apparently general explanatory theory of the *Enquiry* Hume appends a rather long

[79] For the sake of convenience, I will call those mental qualities Hume disapproves of morally ‘useless or monkish virtues.’

[80] In *An Enquiry Concerning Human Understanding* the theory and the counterexample are presented in Section II.

[81] *ES*, 268-280.

text that contains mostly counterexamples to this theory.[82] Any interpretation of Hume's moral philosophy will have to make sense of this rather peculiar procedure.

3.2

Before we can adequately discuss the dilemma in Hume's moral sentimentalism, we need a more detailed account of how Hume explains moral approbation. The dilemma might disappear once we do justice to some of the complexities of Hume's moral sentimentalism. The main question here is this: what exactly is Hume's explanation of how mental qualities that are useful or agreeable either to ourselves or others give rise to the sentiment of moral approbation or blame?

First of all, Hume takes moral approbation of a character (or an action from a character) to be a sentiment that arises only from taking a special point of view:

> 'Tis only when a character is considered in general, without reference to our particular interest, that it causes such a feeling or sentiment, as denominates it morally good or evil. (*T*, 472)

How does a character considered from this point of view cause in us the pleasant or painful sentiments of approbation or blame? In the *Enquiry* Hume answers this question simply by

[82] *AD*, 324-343. Yet another case from Hume's theory of taste: in the essay *Of the Standard of Taste* Hume argues that taste (and the sentiment of beauty that comes from it) is independent of "*bigotry* or *superstition*" (*ES*, 247); but he seems to have no qualms in acknowledging that there is, for example, a taste – obviously *bigoted* – for the works of Paul Bunyan.

pointing to a fact about our nature: “Humanity or a fellow-feeling with others” is “experienced to be a principle of human nature” (*E*, 219n.). Since “no man is absolutely indifferent to the happiness and misery of others” (*E*, 220n.), mental qualities of a person that are useful and agreeable for himself and for others obviously recommend themselves.

In the *Treatise*, Hume goes deeper into the mechanics of our fellow-feeling with others. He explains it by *sympathy*, i.e., our faculty of somehow *converting* an idea of somebody else’s expression of a passion into a (secondary) impression of our own.[83] In the context of morality sympathy works as follows:

> When any quality, or character, has a tendency to the good of mankind, we are pleas’d with it, and approve of it; because it presents the lively idea of pleasure; which idea affects us by sympathy, and is itself a kind of pleasure. (*T*, 580)

(Strictly speaking, the idea of the apparent happiness of other people cannot be a pleasure itself, i.e., an impression. By sympathy the idea is *converted* into a pleasant impression. It is this pleasure that then produces the impression that is, strictly speaking, the impression of moral approbation, which is itself pleasant.)

Thus, it is through the workings of sympathy that perceived useful and agreeable mental qualities give rise to the pleasant (simple) impression of moral approbation. For Hume, moral approbation is, however, not directly linked to sympathy. Usually the operations of sympathy vary with the closeness of our relation to the persons with whom we sympathize. But moral approbation does not seem to behave in this way: “We give the same approbation to the same

[83] See *Treatise*, Book II, Part I, Section XI.

moral qualities in *China* as in *England*. The sympathy varies without a variation in our esteem" (*T*, 581). Hume takes this phenomenon into account by a more counterfactual analysis of moral approbation:

> We blame equally a bad action, which we read of in history, with one perform'd in our neighbourhood t'other day: The meaning of which is, that we know from reflexion, that the former action wou'd excite as strong sentiments of disapprobation as the latter, were it plac'd in the same position. (*T*, 584)

For Hume, a sentiment of praise or blame counts as moral approbation or disapprobation only if it has been corrected by *reflexion*, i.e., if it arises from "some *steady* and *general* points of view" (*T*, 581-82). From such general points of view we have to correct our self-interested sentiments of affection or hatred when we assess the characters of other people. We have to take into account the effects of their qualities not only on our own interests, but also on the interests of everybody else affected by them. Once we take this point of view, once we consider how we would react were we in other people's shoes, sympathy will convert the idea of the generally pleasant consequences of useful and agreeable qualities into impressions of pleasure and approval. This type of approval is *moral* approbation. It is a calm rather than violent passion, universal in its application, not enslaved by the caprices of self-interest, and allows for general (and at least partially rule-governed) systems of praise and blame:

> It [the notion of morals] also implies some sentiment, so universal and comprehensive as to extend to all mankind, and render the actions and conduct, even of the persons most

> remote, an object of applause or censure, according as they agree or disagree with that rule of right which is established. (*E*, 272)

Hume's account of the sentiment of moral approbation is mirrored in his discussion of the sentiment of beauty in the essay *Of the Standard of Taste*.[84] Just as in moral approval I am to feel or judge from a general point of view, in aesthetic[85] approval I am to "forget, if possible, my individual being and my peculiar circumstances" (*ES*, 239) and to disregard the influence of my prejudices and interests. In short, both moral and aesthetic approbation is to come from the point of view of "man in general" (*ES*, 239).[86]

3.3

We are now in a better position to return to the dilemma in Hume's moral sentimentalism. In this section I discuss whether Hume could claim that the approbation of useless or monkish virtues is a type of sentiment different from that of moral approbation. Sections 3.4 to 3.6 examine whether

[84] *ES*, 226-249.

[85] This (convenient) term is, of course, not used by Hume.

[86] Given these similarities in moral and aesthetic approbation, one might ask how Hume distinguishes the sentiment of beauty from the sentiment of virtue. In contrast to the sentiment of moral approbation, the sentiment of beauty appears to be entirely *disinterested*: it does not express a sympathetic interest in general utility and agreeableness. (For some complications concerning the relation between taste, i.e., the sense of beauty, and the moral sense, see *Of the Standard of Taste* (*ES*, 245-247).)

he could consider the approbation of useless or monkish virtues as a *moral*, though *mistaken*, sentiment.

We have seen that for Hume not every sentiment of praise or blame counts as moral approbation. Hume could explain approval of useless or monkish virtues as resulting from a reflective process of correcting original sentiments of praise or blame that differs from a process of *moral* correction of them. Moreover, Hume claims that we have a tendency of confusing distinct sentiments. For instance, sentiments "from interest and morals, are apt to be confounded, and naturally run into each other" (*T*, 472). In a similar way, we might mistake the sentiment of admiration of useless or monkish mental qualities for the sentiment of moral approbation, especially since both types of sentiments seem to be the results of analogous mental operations, i.e., operations of reflection on naturally given sentiments. "But this hinders not, but that the sentiments are, in themselves, distinct; and a man of temper and judgment may preserve himself from these illusions" (*T*, 472). Such a *man of temper and judgment* will recognize that any admiration he might feel for the character traits of someone like Pascal is not really moral approbation. Thus, the apparent dilemma in Hume's moral sentimentalism seems to disappear.

As it stands, this reply is not very satisfactory. Hume needs to explain what exactly distinguishes the sentiment of moral approbation from other sentiments of approval. Otherwise, his talk of our tendency of *confounding sentiments* remains unintelligible. In Book II of the *Treatise*, types of passions, in particular the indirect passions, are distinguished by their respective causes and effects. For Hume, moral sentiments belong to the indirect passions and are thus characterized by their causal context.[87] But this leads to another problem: either Hume

[87] In taking moral sentiments to be indirect passions I follow Árdal's interpretation of Hume's theory of the passions. (See Árdal, ch. 6). This is not the place to defend this interpretation (which allows for a very plausible

can define the sentiment of moral approbation as the (secondary) impression that is caused by the idea of the general utility and agreeableness to mankind of certain mental qualities,[88] or he can define it as the impression which is the effect of any ideas we have when we consider mental qualities from "*steady* and *general* points of view" (*T*, 581-82).

If he takes the first alternative, he can defend his theory of the origin of morals against apparent counterexamples, but only at the price of making it a tautology: that those sentiments which are caused by ideas of the generally useful and agreeable effects of certain mental qualities are caused by precisely those ideas is, though true, unimpressive experimental psychology.

If he takes the second alternative, he is no longer in a position to claim that the approval of useless or monkish virtues is not really moral approbation. Approval of 'dismal' qualities can very well operate from a *steady and general point of view*. If such approval counts as moral approbation, Hume is, of course, wrong about the origin of morals. Thus, unless he can specify the general point of view *appropriate* for moral approbation, he is in no position to denounce seemingly moral approval as an illusion resulting from a confusion of distinct impressions.

Hume sometimes seems to argue that in the approval of useless or monkish virtues an illusion arises not from confounding distinct sentiments, but from mistaken ideas about utility or

reading of Hume's remarks about the relation between the paradigmatic indirect passions of pride and humility and the sentiments of moral approbation and disapprobation). For a contrary view, see Norman Kemp Smith, *The Philosophy of David Hume* (London: Macmillan, 1941), ch. 7.

[88] Caused by this idea and the ensuing impression of pleasure. (More precisely, the sentiment arises from this impression of pleasure and two distinct ideas: the idea Hume calls the 'subject' of the sentiment and the idea he calls its 'object.' Hume claims that indirect passions are characterized by a *double relation* between two ideas and two impression. See *Treatise*, Book II, Part I, Sections IV and V.)

agreeableness; errors about what really is to the benefit of mankind.[89] So, for instance, the "whole train of monkish virtues" (*D*, 226) could seem necessary for reaching the supreme goal of eternal salvation. In Part XII of the *Dialogues Concerning Natural Religion*, Hume argues that such a view depends on errors about the nature of the Deity. To approve of monkish virtues because of their usefulness for avoiding the terrors of eternal damnation involves "both an *absurdity* and an *inconsistency*" (*D*, 226):

> It is an absurdity to believe that the Deity has human passions, and one of the lowest of human passions, a restless appetite for applause. It is an inconsistency to believe, that, since the Deity has this human passion, he has not others also; and, in particular, a disregard to the opinions of creatures so much inferior. (*D*, 226)

Hume could simply stipulate that a sentiment counts as moral approbation only if it is not caused by any such absurd, inconsistent, or superstitious ideas about the conditions of human well-being.[90] In this way he could rescue his theory of the origin of morals: what seems to be a moral

[89] This may look obvious given Hume's discussion of unreasonable passions in the Section *Of the influencing motives of the will* of the *Treatise* (Book II, Part III, Section III). Note, however, that for Hume, properly speaking, only judgments (or ideas) and not passions can be unreasonable (*T*, 416). Moreover, in this Section Hume does not seem to argue that passions with false judgments or ideas in their causal history differ in type from passions caused by true judgments or ideas.

[90] Note that the *inconsistency* in the passage quoted cannot mean a demonstrable falsehood (*demonstrable* in Hume's sense of 'demonstration.')

sentiment and a counterexample to the theory is nothing but the emotive outgrowth of some absurd or superstitious idea.[91]

To make this argument plausible Hume would have to clarify what distinguishes superstitious and absurd ideas from simply false ones. It would be very odd for Hume to maintain that only sentiments of approval caused by *true* ideas of general usefulness and agreeableness should count as moral sentiments. In his general account of moral approbation he emphasizes that sympathy operates on ideas obtained from fallible inferences:

> No passion of another discovers itself immediately to the mind. We are only sensible of its causes or effects. From *these* we infer the passion: And consequently *these* give rise to our sympathy. (*T*, 576).

Thus, if sympathy produces moral approbation in the way Hume describes, some moral sentiments will be caused by false ideas about utility and agreeableness. Once this is admitted, it seems most implausible to argue that one particular class of such false ideas produces sentiments of approval distinct in type from sentiments of *moral* approbation. (Moreover, I will show in Section 3.6 that Hume does not think that an absence of (religious) superstition would guarantee an approbation of only those qualities he calls ‘useful and agreeable.’)

So far no plausible way out of the dilemma in Hume’s moral sentimentalism has emerged. The apparent counterexamples to his theory of the origin of morals – sentiments of approbation of useless or monkish virtues – still seem to refute it. We have seen that Hume could rescue it by making it true by definition: only the approbation of generally useful and agreeable

[91] After all, according to Hume, all idea have a tendency to give rise to (secondary) impressions (*T*, 373).

mental qualities *counts* as moral approbation. This might not be a move Hume would want to avoid.[92] If he can provide good reasons for why we should approve of mental qualities from the point of view of utility and agreeableness, making his apparently 'experimental' account of the origin of the sentiments of moral praise or blame trivially true might not bother him.[93]

Up to this point we have only considered what Hume has to say about the nature of the sentiment of moral approbation and its origin. However, both in the *Treatise* and in the *Enquiry*, Hume does not clearly separate discussing what constitutes the sentiment of moral approbation from arguing why we should cultivate it. His real concern might not be an explanation of the "true origin of morals" (*E*, 173) in terms of universal psychological principles, but a defense of a particular point of view of approbation and disapprobation. Hume's views about how his own version of such a point of view recommends itself might shed more light on his arguments against advocates of useless or monkish virtues. The dilemma in Hume's moral sentimentalism could be a pseudo-problem that arises only from taking too literally his pronouncements about following the *experimental method*. Hume's real interest might not be in the causes of any given sentiments of moral approbation. Explaining the sentiments of advocates of useless or monkish virtues might not really matter to him, at all. This, of course, raises the question of how seriously

[92] In his essay *Of the Standard of Taste*, Hume calls any seeming sentiment of beauty that is influenced by what he is pleased to call "*bigotry* or *superstition*" (*ES*, 247) "erroneous" and "perverted" (*ES*, 241). It is clear, however, that what he means by this is that the sentiment is not really a sentiment of *beauty*.

[93] Given his dislike for verbal disputes, Hume would, however, hardly think that simply *defining* moral approbation as a sentiment of approval with a particular type of causal history cuts any philosophical ice. Moral approbation, defined in this way, might still be some kind of error.

we should take Hume’s account or moral sentiments as an application of the experimental method. His ‘psychology’ might prove to be of a rather special kind.

3.4

As we have seen, for Hume moral approbation occurs from “*general* and *steady* points of view” (*T,* 581-82). He argues that taking such points of view is necessary for the practice of moral assessment:

> Our situation, with regard both to persons and things, is in continual fluctuation; and a man, that lies at a distance from us, may, in a little time, become a familiar acquaintance. Besides, every particular man has a peculiar position with regard to others; and ‘tis impossible we cou’d ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view. (*T*, 581)

Without a general system of praise and blame we could not form the steady moral sentiments that give meaning to the general terms of moral language which allow us to communicate our sentiments of praise or blame. In the *Enquiry*, Hume makes a similar point, but emphasizes that the utility of such a system goes beyond facilitating conversation:[94]

[94] In the eighteenth century ‘conversation’ could have a wider meaning than today. In the quoted text from the *Treatise*, however, Hume seems to use it in the modern sense. In the parallel passage in the *Enquiry* he seems to use

> General language, therefore, being formed for general use, must be moulded on some more general views, and must affix the epithets of praise and blame, in conformity to sentiments, which arise from the general interests of the community. (*E*, 228)

Thus, the interests of the community seem to justify taking a moral point of view. The calm passions of the moral point of view may keep in check the violent passions issuing in purely self-interested and short-sighted actions that could prove pernicious to the welfare of society.

Hume’s argument here can show at best the need for taking *some* steady and general point of view. But, as we have already pointed out, the approbation of useless or monkish virtues could very well operate from general viewpoints. Hume may, of course, reply that such viewpoints do not take into account what he calls ‘the general interests of the community.’ But now he needs to explain what these interests are and why useless or monkish virtues cannot be in anyone’s or any community’s interest.

In the next Section we shall briefly examine Hume’s notion of what is in our interest. To this end we must consider his account of those mental qualities or dispositions that he believes are necessary for our happiness. If Hume could show that the approbation of useless or monkish virtues is the result of dispositions unfavorable to the universally pursued end of happiness, he could easily denounce them as imprudent.

‘conversation’ and ‘discourse’ interchangeably. For ‘conversation’ in the wider sense his term appears to be ‘social intercourse’ (*E*, 228-29).

3.5

In the *Enquiry* Hume proposes the following thought experiment:

> Let a man suppose that he has full power of modeling his own disposition, and let him deliberate what appetite or desire he would choose for the foundation of his happiness and enjoyment. (*E*, 281)

Let us suppose that in modeling his own disposition the man also chooses his standards of approbation. In the essay *Of the Delicacy of Taste and Passion* Hume actually discusses such a choice between two dispositions: delicacy of passion and delicacy of taste. He argues that "everyone will agree with me, that […] delicacy of taste is as much to be desired and cultivated as delicacy of passion is to be lamented, and to be remedied, if possible" (*ES*, 5). The argument against the disposition of the person with a delicacy of passion is important here because this person lets herself be guided by passions like violent love and hatred rather than by the calm passions of morality (as understood by Hume).

Hume gives two reasons for his rejection of delicacy of passion: First, "every wise man will endeavour to place his happiness on such objects chiefly as depend on himself: and *that* is not to be *attained* to much by any other means as by this delicacy of sentiment [delicacy of taste]" (*ES*, 5). The disposition of delicacy of passion puts our happiness at the mercy of conditions beyond our control, i.e., the conditions under which (violent) passions are inflamed. For someone with delicacy of taste, happiness depends less on fortuitous circumstances and the caprices of others: "The good or ill accidents of life are very little at our disposal; but we are pretty much masters what books we shall read, what diversions we shall partake of, and what company we shall keep" (*ES*, 5).

Second, delicacy of taste is "favourable to love and friendship" (*ES*, 7). It "improves the temper" and gives rise to emotions that "cherish reflection; dispose to tranquility; and produce an agreeable melancholy, which, of all disposition of the mind, is best suited to love and friendship" (*ES*, 6-7). Moreover: "One that has well digested his knowledge of both books and men, has little enjoyment but in the company of a few select companions" (*ES*, 7).

From these considerations Hume appears to conclude that the life of the person with a delicacy of taste is happier than that of the person with a delicacy of passion. The life of the latter is likely to be unhappy most of the time since his "sensibility of temper" is such that when he "meets with any misfortune, his sorrow or resentment takes entire possession of him, and deprives him of all relish in common occurrences of life; the right enjoyment of which forms the chief part of our happiness" (*ES*, 4). Here Hume seems to presuppose a notion of happiness that for anyone (no matter what his actual passions and wishes happen to be) would justify the same judgments about the respective desirability of dispositions.

Similarly, in the essay *The Sceptic*, after pointing out that it "must be obvious to the most careless reasoner, that all dispositions of mind are not alike favourable to happiness" (*ES*, 168), Hume lists conditions for the "happiest disposition of mind."[95] The passions of the happiest disposition must be:

[95] In this essay Hume argues against the conception of happiness advocated by the speakers in the essays *The Epicurean*, *The Stoic*, and *The Platonist* (*ES*, 138-58). I take it that in the *The Sceptic* Hume speaks in his own voice. In defense of this interpretation it must suffice here to point out the tone and diction of the essay is close to that of the *Enquiry* and markedly different from the tone of parody in the three preceding essays. That *The Sceptic* can be

> (i) neither too violent nor too remiss;
>
> (ii) benign and social; not rough and fierce;
>
> (iii) cheerful and gay; not gloomy and melancholy;
>
> (iv) such that the enjoyment of their objects is steady and constant and conveys durable pleasure and satisfaction. (*ES*, 167)[96]

Hume claims that satisfaction of the first three requirements guarantees that our passions are as "agreeable to the feeling" (*ES*, 167) as possible. By the fourth condition he intends to exclude the life of pleasure and the life of business in pursuit of "external objects" (*ES*, 167) as forms of true happiness. The life of pleasure leads to "satiety and disgust" (*ES*, 168) and cannot provide durable satisfaction. The life of business puts us at the mercy of ever changing external circumstances that prevent steady and constant enjoyment. Hume concludes, rather abruptly, that the *virtuous* disposition best meets these requirements. This disposition

> leads to action and employment, renders us sensible to the social passions, steels the heart against the assaults of fortune, reduces the affections to a just moderation, makes our own

used to illuminate Hume's own moral theory is accepted in most of the Hume interpretations that emphasize the unity of his works: see, for instance, Annette Baier's *A Progress of Sentiments* (Cambridge, MA: Harvard University Press, 1991), ch. 8. For a similar, though slightly qualified view, see Donald T. Siebert, *The Moral Animus of David Hume* (London and Toronto: Associated Universities Press, 1990), 187-194.

[96] I have extracted these criteria from three brief paragraphs in Hume's text.

> thoughts an entertainment to us, and inclines us rather to the pleasures of society and conversation, than to those of the senses. (*ES*, 168)

Someone with this disposition will show “a steady preference for the greater happiness” (*E*, 279) and approve of mental qualities which are useful and agreeable to himself and others.

In the famous discussion of ultimate ends in Section V of the Appendix I to the *Enquiry*, Hume claims that “something must be desirable on its own account, and because of its immediate accord or agreement with human sentiment and affection” (*E*, 293). He then adds that virtue, i.e., the life of the virtuous disposition, is desirable on its own account since it agrees immediately with our sentiment. This immediate agreement is the natural standard of virtue and vice, and it is given by our “internal frame and constitution” (*E*, 294). In other words, Hume seems committed to general principles of (immediate) approbation of dispositions, principles which are uniform in human beings.[97] Note that only principles of *immediate* approbation are claimed to be universal. When we corrupt our internal frame and constitution by the artificial arguments of philosophy and religion, the natural standard of morality will no longer have a hold on us. If, like Pascal or the ‘Epicurean,’ ‘Stoic,’ or ‘Platonist,’ we “affect […] *artificial* lives” (*AD*, 343), the operations of the natural principles of approbation are disrupted by the artifices of superstition or philosophical enthusiasm.

Hume’s argument against the advocates of useless or monkish virtues could now be read as claiming not that they are mistaken about the nature of their own sentiments of approval (by

[97] See also *Of the Standard of Taste* (*ES*, 233) for the parallel case of the standard of taste. (If Hume had been given to mythic representation, he might have ascribed natural sentiments of moral or aesthetic approbation to a *state of nature*.)

somehow confounding them with other sentiments), but that they are mistaken in their fundamental choice of a disposition and in their following of a standard of moral approbation that differs from the *natural* standard. Although their sentiments of praise and blame can count as sentiments of moral approbation or disapprobation, they are the result of an erroneous conception of happiness. Since everyone wants to have a disposition that makes him happy,[98] having monkish and useless dispositions and corresponding moral sentiments is simply self-defeating.

3.6

Interpreted in the way of the preceding section, Hume faces another problem: he holds that there is a natural standard for the evaluation of dispositions which, under certain conditions, everybody would endorse and apply. But for just about any standard of evaluation, counterfactual conditions can be specified under which everybody would follow it. Hume probably would not consider this a problem since the distinctive feature of the standard he is pleased to call the 'natural standard' is precisely that it would be followed under *natural* conditions: the conditions under which we approve or disapprove immediately, without any intermediary artifices. But now Hume needs to answer the question of why we should follow this 'natural' standard rather than some 'artificial' one.[99]

[98] "No man would ever be unhappy, could he alter his feelings" (*ES*, 168).

[99] In his theory of taste Hume also singles out "natural sentiments [of beauty]" (*ES*, 241) that accord with a natural standard of taste.

Pascal, the ‘Epicurean,’ the ‘Stoic,’ and the ‘Platonist’ might all very well agree with Hume that if they gave up their religious or philosophical beliefs, they would have sentiments of moral approbation only for useful and agreeable qualities (in accordance with the ‘natural’ standard of that which Hume calls ‘virtue’). But this counterfactual truth alone gives them no reason to switch from their ‘artificial’ standard to the ‘natural’ one, not even a reason to wish for such a switch. They also would reject Hume’s further claim that only the disposition we would approve of immediately or ‘naturally’ can lead to happiness. After all, their ‘artificial’ standards presuppose conceptions of happiness that differ from Hume’s. For instance, that an Epicurean disposition does not lead to pleasures as durable as the disposition of Hume’s ‘virtuous man’ is of little concern to the ‘Epicurean.’ As the ‘Epicurean’ tells Caelia, the “mistress of his wishes” (*ES*, 145): “Consider rather, that if life be frail, if youth be transitory, we should well employ the present moment, and lose no part of so perishable an existence” (*ES*, 145).[100] Similarly, to the person with a delicacy of passion the happiness from a delicacy of taste will appear only insipid.[101]

Hume and the advocates of monkish and useless virtues do not share a common notion of happiness that would allow adjudicating the question of what is the standard of virtue and vice. Given his own conception of happiness and virtue, Hume cannot argue that the ‘Epicurean’ will be happier *in Epicurean terms* by becoming ‘virtuous’: the happiness from the passionate pleasures of the moment is not commensurable with that from the durable pleasures of Hume’s

[100] We need not decide here whether Hume’s Epicurean (*galant* in the style of the 18th century, and perhaps something of a libertine) is a faithful disciple of *Epicurus*.

[101] Remember the “agreeable melancholy” (*ES*, 7) produced by delicacy of taste. We might also imagine what *Pascal* would say about a delicacy that finds its happiness in pleasant *diversions* (*ES*, 5).

man of taste. Analogous points can be made, of course, about comparing the respective happiness produced by Hume’s favorite disposition and the dispositions of Pascal, the ‘Stoic,’ or the ‘Platonist.’

Hume seems to recognize the problem when he considers a person “of so perverse a frame of mind, of so callous and insensible a disposition, as to have no relish for virtue” (*ES*, 169). This person replies to Hume’s commendation of the pleasures of the ‘virtuous’ disposition by saying – in Hume’s own words – that “theses were, perhaps, pleasures to such as were susceptible of them; but that, for his part, he finds himself of a quite different turn and disposition” (*ES*, 169-70). To this Hume answers as follows:

> My philosophy affords no remedy in such a case, nor could I do any thing but lament this person’s unhappy condition. But then I ask, If any other philosophy can afford a remedy; or if it be possible, by any system, to render all mankind virtuous, however perverse may be their natural frame of mind? (*ES*, 170).

This passage is remarkable in many ways. First, the talk of ‘this person’s unhappy condition,’ her ‘perverse frame of mind,’ and her lack of ‘virtue’ amounts to nothing more than Hume’s acknowledgment of a the fact that she has a different disposition, and to his expression of a sentiment of disapprobation of her frame of mind. This disapprobation is entirely from the point of view of Hume’s own standard and without any claim to a hold on the ‘perverse’ other person. Here, Hume no longer insists that about the preferability of disposition “every one will agree with him” (*ES*, 5).

Second, note that other persons' 'perverse frames of mind' are also called 'natural': 'natural' must here be read as opposed to 'artificial.' It can hardly be opposed to 'unusual' or 'miraculous': in the context of the passage it would be pointless for Hume to emphasize the frequent occurrence of 'perverse frames of mind.' (That such frames of mind are no miracles goes without saying.) This means that Hume can no longer use the '*natural* standard' of virtue and vice to support his preferred disposition. Now, the 'natural standard' seems to be not really a standard: it appears to sanction both 'virtuous' and 'vicious' dispositions since Hume acknowledges that some persons 'naturally' ('immediately') approve of dispositions less than 'virtuous' (in *Hume's* sense of this term, of course).

Third, it is easy to agree with Hume that the arguments of his or any other's moral philosophy will not 'render all mankind virtuous.' But the significance of this truth is unclear. Obviously we cannot expect a moral philosophy with a conception of virtue that differs from Hume's to promote those mental qualities Hume himself calls 'virtuous.' Furthermore, none of Hume's opponents believes that the success of moral philosophy depends on its power to 'render all mankind virtuous.' But, like Pascal, the 'Epicurean,' 'Stoic,' and 'Platonist,' Hume seems to be engaged in answering the question of which disposition really *is* "to be desired and cultivated," and which "to be lamented, and to be remedied, if possible" (*ES*, 5). That the answers of moral philosophy will not change the disposition of all mankind does not imply they cannot be based on a universal standard. Hume, of course, often expresses his agreement with this, but he seems unable to justify that such a universal standard exists and is to be followed.

Thus in disputes about the standard of virtue and vice, Hume's position *vis-à-vis*, for instance, that of the 'Epicurean' seems to be no better than that of the 'Epicurean' *vis-à-vis* Hume. If Hume accuses the 'Epicurean' of being overcome by a philosophical system or enthusiasm (or an overemphasis on the frailty of life), there appears to be no reason why the latter could not simply return the charge: Hume himself is in the grip of a philosophical system in favor of a 'natural' standard of virtue and vice, a standard that proves only imaginary.

3.7

The discussion of the dilemma in Hume' moral sentimentalism has led us to his remarks in defense of a particular frame of mind (i.e., the 'virtuous' disposition) which entails moral approbation from the point of view of general utility and agreeableness. It now appears, however, that in Hume's moral philosophy we can find neither a general causal account of the origin of the peculiar sentiments of moral approbation and disapprobation nor a justification of a particular standard of virtue and vice. Hume's 'experimental' moral psychology seems to be only a theory of the moral sentiments of the 'virtuous' person. Although Hume claims that this disposition agrees with the 'natural' standard of virtue and vice, he does not provide a defense of the universality of this standard.

Such a conclusion should make us reconsider what the aim of Hume's moral philosophy could be. In the last Section of Book I of the *Treatise* Hume describes the general motive for his philosophical investigations: a natural inclination 'to carry my view into all those subjects, about which I have met with so many disputes in the course of my reading and conversation" (*T*, 270). With regard to the questions of moral philosophy he tells us: "I cannot forbear having a curiosity to be acquainted with the principles of moral good and evil. I am uneasy to think I approve of one object, and disapprove of another […] without knowing upon what principles I proceed" (*T*, 270-71). In writing the *Treatise*, the *Enquiry*, and the *Essays*, Hume certainly dispels this uneasiness about not knowing the principles of his sentiments of moral praise and blame. The question, then, is this: in what sense does this exercise in philosophical autobiography[102] amount to a defense of these principles?

In the *Conclusion* of the *Treatise*, Hume addresses the “lovers of virtue” (*T*, 619), i.e., the persons with the ‘virtuous’ disposition, and presents them with one last reflection:

> It requires but very little knowledge of human affairs to perceive, that a sense of morals is a principle inherent in the soul, and one of the most powerful that enters into the composition. But this sense must certainly acquire new force, when reflecting on itself, it approves of those principles, from whence it is deriv’d, and finds nothing but what is great and good in its rise and origin. (*T*, 619)

The sense of morals of the ‘virtuous’ person arises from a “noble source”: “an extensive sympathy with mankind” (*T*, 619). Thus the ‘virtuous’ person approves not only of ‘virtue,’ but also of the principles from which it is derived.[103] So, Hume could be read as providing a *positive* genealogy of a particular type of morality, a genealogy that has the practical purpose of reinforcing the morality whose ‘origin’ it describes.[104]

[102] This autobiography shows Hume, of course, not as a solitary moral judge, but as a member of a society in which moral principles serve useful purposes “in company, in the pulpit, on the theatre, and in the schools” (*T*, 603).

[103] ‘Principle’ refers here, of course, not to an abstract moral rule, but to the origin of the moral sentiment (in the character of the virtuous person).

[104] In this way Hume’s genealogy differs from the famous *Genealogy of Morals*, a *negative* genealogy in which Nietzsche attempts to loosen the hold certain moral notions have on us.

If we read Hume in this way, the starting point of this essay, the apparent dilemma in Hume's moral sentimentalism, no longer presents a problem. The seeming counterexamples to his account of moral approbation are only counterexamples to a theory with a much wider scope than his own account of the 'origin of morals.' Hume does not aim at an explanation in terms of the universal psychological laws of any sentiment of moral approbation, but restricts himself to making sense of the frame of mind of the 'virtuous' person. In fact, it is important for Hume to maintain that the explanation of the approbation of useless or monkish virtues reveals principles that differ from those that explain the approbation of useful and agreeable mental qualities. Hume seems to have no doubt that a genealogy of useless or monkish dispositions would bring to light most unsavory 'origins.' That these other dispositions present to Hume such origins prevents him from seriously considering them as alternatives to his own 'virtuous' disposition.

We can still see Hume's account of 'the origin of morals' as an application of the 'experimental method.' Starting with the fact of the existence of the 'virtuous' disposition and its corresponding mode of moral approbation, Hume derives from experience the principles governing this kind of moral approbation. In accordance with his methodological principles he so "deduces general maxims from a comparison of particular instances" (*E*, 174). Again, the scope of these experimentally derived general maxims is limited: "An experiment, which succeeds in the air, will not always succeed in a vacuum" (*AD*, 343). Hume's 'experiment' about the principles of moral approbation is designed only for a special case: the moral sentiments of the 'virtuous' person.[105]

Hume would not deny that the defenders of useless or monkish virtues could construct – in their own terms – a positive genealogy of their morality which would reinforce their

[105] Or, less misleadingly, the 'tolerably virtuous person.' See the passage quoted at the end of this Section.

adherence to it.[106] His own genealogical defense of the ‘virtuous’ disposition is purely internal. It does not aim at convincing the ‘non-virtuous’ of the error of their ways. Hume, after all, does not believe that the arguments of moral philosophy have the power of changing dispositions: “Whoever considers, without prejudice, the course of human actions, will find, that mankind are almost entirely guided by constitution and temper, and that general maxims have little influence” (*ES*, 169).[107] Although he states in the *Enquiry* that the “the end of all moral speculations is to teach us our duty” (*E*, 172) and, in the *Essays*, discusses methods of “correcting the temper” and “reforming the mind” (*ES*, 169-170), he emphasizes the limits of moral instruction and reform:

> Where one is thoroughly convinced that the virtuous course of life is preferable; if he have but resolution enough, for some time, to impose a violence on himself; his reformation needs not to be despaired of. The misfortune is, that this conviction and this resolution never can have place, unless a man be, before-hand, tolerably virtuous. (*ES*, 171)

[106] They could, of course, also give a *negative* genealogy of the Humean ‘virtuous’ disposition. See, for instance, Alasdair MacIntyre, *Whose Justice? Whose Rationality?* (Notre Dame, IN: University of Notre Dame Press, 1988), chs. 15 and 16. (See also Dorothea Krook, *Three Traditions of Moral Thought* (Cambridge: Cambridge University Press, 1959).)

[107] In the terms of Hume’s friend Denis Diderot, we come to the ‘virtuous’ disposition by being *happily born*. If we are so lucky our “natural frame of mind” (literally the frame we are born with) is not “perverse” (*ES*, 170).

3.8

In her book *A Progress of Sentiments*, Annette Baier interprets Hume's moral philosophy as a reflexive genealogy of the kind outlined in the preceding Section. She claims that for Hume "reflexive self-approval" is "the perfection of practical reason,"[108] and that a disposition of moral approbation passes the "test of reflexivity" if it is capable of "bearing its own survey."[109] With regard to the capacity of Hume's 'virtuous' disposition to bear its own survey, Baier points out that "the circularity of appealing to the survey and judgment of cheerful, friendly, wit-loving people [i.e., people with the 'virtuous' disposition and with delicacy of taste] to get the approval of cheerfulness, friendliness, and wit seems to be taken [by Hume] to be a wholly virtuous circularity."[110]

But the same circularity in applying the test of reflexivity might, of course, appear in 'testing' the disposition of the person with a delicacy of passion, the dispositions of the 'Epicurean,' the 'Stoic,' and the 'Platonist,' and perhaps even in testing Pascal's monkish disposition. The proposed test of reflexivity might not rule out any of the dispositions of approbation Hume himself considers.

If the test of reflexivity is interpreted as ruling out sentiments that are in some sense self-defeating, Hume might have an argument against the approbation of monkish virtues. As Baier writes: "The sour may well approve of sourness, the hard-hearted of ruthlessness […] but the self-hating cannot coherently love themselves for their self-hatred. […] Humility as a virtue

[108] Annette Baier, *A Progress of Sentiments* (Cambridge, MA: Harvard University Press, 1991), 277.

[109] Baier, 215-217. 'Bearing its own survey' is a phrase of Hume's (*T*, 620).

[110] Baier, 217.

faces a paradox, namely that the very approval of it seems to threaten to destroy the thing approved."[111] (We may, however, wonder whether the friends of humility would be impressed by the fact that humility cannot pass the 'test of reflexivity.' It might seem to be the very essence of real humility *not* to reflect on itself, *not* to take its own survey. Moralists for whom humility is a central virtue have typically considered reflection a source of moral corruption, and have championed *simplicity* over reflexivity.[112]) But even if we grant that the approbation of monkish virtues somehow undermines itself, it seems quite clear that this problem is not to be found in most of the other perspectives on moral approbation that rival Hume's own. Furthermore, Hume would have to show why an incoherence in the approval of certain virtues is really undesirable, He could describe such an incoherence only as a succession of distinct passions that in some way creates turmoil in the mind. But given our tendency to "tiresome indolence" (*T*, 452), this might be all too welcome.

We must now ask how reflexive self-approval is to be distinguished from complacent self-congratulation, and whether Hume's genealogy of morals is more than a scheme of self-deception. Is it more than just the least inconvenient way of relieving an uneasiness about the

[111] Baier, 215-16.

[112] One of the moralists most read in the 18th century, namely Fénelon (with whose writings Hume was familiar), never tires of warning of the dangers of reflection. Similarly, Hume's great philosophical antagonist, Jean-Jacques Rousseau, will later claim in the *Discourse on the Origin of Inequality* that "the state of reflection is a state against nature" (*Oeuvres completes*, t. III, ed. Marcel Raymond et Bernard Gagnebin (Paris: Gallimard, 1964), 138). It is also be worth pointing out that by claiming that for Hume reflexivity is the 'perfection of practical reason' Annette Baier gives Hume's moral sentimentalism a somewhat surprising intellectualist turn. (For some contemporary variations on this theme, see Julia Driver, *Uneasy Virtue* (Cambridge: Cambridge University Press, 2001).

principles of virtue and vice, and of putting scruples about their justification (not to mention scruples of conscience) to rest?

Given his views about the limited powers of philosophy, it seems that Hume would hardly be disturbed by these questions. He would characterize his own enterprise in moral philosophy as the natural effect of the passion he calls "curiosity or the love of truth" (*T*, 448), the passion that he claims to be the motive of all philosophy. In his discussion of this 'love of truth' (which he compares to the passion for *hunting*[113]) he argues that it can be satisfied only by endeavors that are difficult and appear important and useful. It should not surprise us that his moral philosophy meets these criteria: we do not "come to the knowledge of it without difficulty, and without any stretch of thought and judgment" (*T*, 449). It is also attended with an "idea of utility" (*T*, 449) that allows us to imagine we are engaged in a pursuit advantageous to the interest of mankind: "What philosophical truths can be more advantageous to society, than those here delivered?" (*E*, 279).[114] Moreover, this idea of utility associated with the theory guarantees its stability (even independently of its other merits): "Truths which are *pernicious* to society, if any such there be, will yield to errors which are salutary and *advantageous*" (*E*, 279).

Whether all this amounts to a *reductio ad absurdum* of philosophical reflection on morality, we need not decide here. It might be amusing to end – on the theme of reflexive self-approval – with the account of a small episode from the autobiography of a good friend of Hume's:

[113] See *T*, 451-452.

[114] This remark is, of course, ironic since Hume believes philosophical truths to have very little influence in common life and society.

> In my first Voyage from Boston, being becalm’d of Block Island, our People set about catching Cod & hawl’d up a great many. Hitherto I had stuck to my Resolution of not eating animal Food; and on this Occasion, I consider’d with my Master Tryon, the taking every Fish as a kind of unprovok’d Murder, since none of them had or ever could do us any Injury that might justify the slaughter. – All this seem’d very reasonable. – But I had formerly been a great Lover of Fish, & when this came hot out of the Frying Pan, it smelt admirably well. I balanc’d some time between Principle & Inclination: till I recollected, that when the Fish were opened, I saw smaller Fish taken out of their Stomachs. – Then, thought I, if you eat one another, I don’t see why we mayn’t eat you. So I din’d upon Cod very heartily and continu’d to eat with other People, returning only now & then occasionally to a vegetable Diet. So convenient a thing it is to be a *reasonable Creature*, since it enables one to find or make a reason for every thing one has a mind to do.[115]

[115] Benjamin Franklin, *The Autobiography* (New York: First Vintage Books/The Library of America, 1990), 35.

4.0 SUBJECTIVISM, UTILITY, AND AUTONOMY

Subjectivists in ethics have dismissed Kantian conceptions of practical rationality for a variety of reasons.[116] Contrary to Kant, subjectivists claim that all reasons for action depend on our desires, pro-attitudes, preferences, or some unspecified psychological amalgam called the 'motivational set.'[117] Moral reasons, in particular, are then seen as contingent on a subclass of these entities.

In order to elaborate this conception of reasons for action, many subjectivists have found it convenient to use the conceptual resources of Rational Choice Theory.[118] With its concepts of preference and utility, Rational Choice Theory promises to give precise content to the notion of acting on the basis of one's desires. It thereby appears to make possible an elegant development of a subjectivist theory of value. Such a theory of value seems to owe its appeal to the apparently unproblematic notion of action for the sake of desire satisfaction or – in the terms of Rational

[116] For representative and influential arguments see J.L. Mackie, *Ethics – Inventing Right and Wrong* (Harmondsworth: Penguin, 1977), ch. 1; Gilbert Harman, *The Nature of Morality* (Oxford: Oxford University Press, 1977), ch. 11; Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press, 1985), ch. 4.

[117] This is Bernard Williams' term in "Internal and External Reasons" (in his *Moral Luck* (Cambridge: Cambridge University Press, 1981).

[118] See, for example, Richard B. Brandt, *A Theory of the Right and the Good* (Oxford: Clarendon Press, 1979); David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986); John C. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge: Cambridge University Press, 1977).

Choice Theory – of acting for the maximization of subjective expected utility. Thus, subjectivism would not commit us to the seemingly extravagant metaphysical claims about human agency found in one of its main rivals, Kant's practical philosophy.

David Gauthier has recently outlined an argument for a *subjectivist* account of value that relies on a conception of the rational agent as capable of a kind of autonomy: an autonomy in the unification of given desires into a coherent whole.[119] According to Gauthier, Rational Choice Theory – by imposing conditions of coherence on our desires – formulates the very conditions under which we can conceive ourselves as autonomous rational agents. Gauthier's argument is an attempt to take up the Kantian concern with autonomy and to enlist it in a defense of subjectivism. For if conceiving ourselves as autonomous rational agents demands acting in accordance with Rational Choice Theory, subjectivism about value will turn out to be justified even from a (broadly understood or, in Gauthier's terms, 'naturalized') Kantian perspective.

I shall argue in this chapter that Gauthier's argument cannot succeed and that this points to a general problem about the relation between subjectivism about value and Rational Choice Theory. In Section 4.1 I sketch in more detail how Gauthier explains the connection between the notion of an autonomous rational agent and the concepts and principles of Rational Choice Theory. This will make it necessary to distinguish Gauthier's notion of autonomy from that of Kant (Section 4.2). In Section 4.3 I argue that by its very nature of being a theory of utility, Rational Choice Theory cannot give content to the notion of autonomous agency as Gauthier understands it. I try to establish this by arguing that we cannot expect determinate practical principles from Rational Choice Theory. Sections 4.4 and 4.5 present two consequences of this

[119] David Gauthier, "The Unity of Reason: A Subversive Interpretation of Kant," in *Moral Dealing* (Ithaca: Cornell University Press, 1990), 110-126.

result: first, that Gauthier’s conception of autonomy is unstable and directs us to Kant’s conception of it (Section 4.4), second, that there seems to be no important role for Rational Choice Theory even in subjectivist accounts of value (Section 4.5).

4.1

In the introduction to his collection of essays *Moral Dealing*, Gauthier comments on the foundation of his maximizing conception of practical rationality:

> I find myself increasingly persuaded by a view of rationality that might be part of a naturalized Kantianism. […] I found myself focusing on the Kantian understanding of reason as unifying our beliefs, desires, and feelings into the experience of a single self – an individual. And this, it now seems to me, provides the deep basis of the maximizing conception of rationality.[120]

Gauthier’s essay “The Unity of Reason: A Subversive Reinterpretation of Kant”[121] sketches such a ‘naturalized Kantianism.’ Gauthier sees an unwarranted asymmetry in Kant’s accounts of the activities of the understanding and the will. The understanding – by means of its pure concepts, the categories – synthesizes the manifold of intuition and thereby makes knowledge possible, and – by the principles of pure understanding – gives *a priori* laws to all appearances. To the

[120] Gauthier, *Moral Dealing* (Ithaca: Cornell University Press, 1990), 7.

[121] Gauthier, *Moral Dealing*, essay 5.

manifold of intuition there exists a corresponding manifold of desire. As the manifold of intuition provides the material for our knowledge, the manifold of desire provides the material for our choice. But the manifold of desire cannot give rise to choice and action without the synthesizing activity of the will. This unifying activity could be seen as guided by ‘pure concepts of the will’:[122]

> For choice to be possible, the desires of the actor must be unified in such a way that they determine a single alternative from those possible actions that are available to her. The actor's desires must be so related that they determine a preferential ordering of the set of alternative possible actions, from which she may then select a maximal element. The familiar ideas of the theory of rational choice correspond to the pure concepts of the will.[123]

The concepts and principles of Rational Choice Theory specify what it means to maximize the satisfaction of one’s preferences. Thus, they might be interpreted as giving a precise content to Kant’s own notion of happiness as the “satisfaction of all our desires.”[124]

[122] Dieter Henrich argues that in his early critical period Kant himself attempted to exploit the analogy between pure concepts of the understanding applying to the manifold of intuition and pure concepts (of choice) applying to the manifold of desire. See Henrich’s “Der Begriff der Sittlichen Einsicht und Kants Lehre vom Faktum der Vernunft,” in G. Prauss, ed., *Kant: Zur Deutung seiner Theorie von Erkennen und Handeln* (Köln: Kiepenheuer & Witsch, 1973). See also Henry Allison, *Kant’s Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 66-70.

[123] Gauthier, *Moral Dealing*, 115-116.

[124] Immanuel Kant, *Critique of Pure Reason*, tr. Norman Kemp Smith (London: Macmillan, 1929), B 834.

Gauthier interprets Kant as claiming that the pursuit of happiness is not guided by the will (i.e., by practical reason), but by natural necessity. Kant does indeed assert that happiness is "an unavoidable determinant of the faculty of desire,"[125] and this Gauthier takes to mean that "we do seek happiness, but we do not *choose* to seek happiness."[126] As Gauthier now argues, this explains why Kant sees no practical role for reason with regard to happiness.

Gauthier's argument here is as follows: Kant wants to find a practical law, i.e., a universal determination of the will that is valid for each rational agent.[127] Now, according to Kant, laws are synthetic and necessary principles; practical laws are synthetic and necessary principles that confront us finite rational beings as imperatives, i.e., they prescribe actions or ends as rationally necessary whether or not we actually will them. But for Kant, practical laws and the pursuit of happiness do not fit:

> Now Kant seems to suppose that a principle prescribing happiness as an end would be synthetic and necessary but not practical, whereas a principle prescribing some action as a means to happiness, if it were necessary, would be practical but not synthetic. Thus happiness cannot give rise to a practical law.[128]

[125] Immanuel Kant, *Critique of Practical Reason*, tr. Lewis White Beck (Indianapolis: Bobbs-Merrill, 1956), Academy 25.

[126] Gauthier, *Moral Dealing*, 114.

[127] Kant, *Critique of Practical Reason*, Academy 19.

[128] Gauthier, *Moral Dealing*, 114.

In other words, if happiness is pursued by natural necessity, no principle *prescribing* happiness as an *end* can determine the will, i.e., be practical.[129] Principles prescribing *means* to happiness reduce to the principle "Whoever wills the end, wills also the means that are indispensably necessary to his actions and that lie in his power,"[130] which, even if it were practical, could not be a law since it is analytic.

It is now a central step in Gauthier's reasoning that Kant's notion of happiness as the satisfaction of all desires makes it impossible to conceive of happiness as an end given by natural necessity. The given manifold of desire in itself is unordered and must be unified before we can aim at an overarching end like happiness or a "maximum of well-being."[131] This unification of the manifold of desire according to principles is then to be understood as the activity of practical reason. In this way happiness *can* give rise to practical laws, namely those principles that are necessary for the unification of the manifold of desire and through this unification allow for a choice that has happiness as its end. Such principles are not analytic but synthetic; and, as already mentioned, Gauthier thinks that Rational Choice Theory provides these principles:

[129] Kant writes: "a command that everyone should seek to make himself happy would be foolish, for no one commands another to do what he already invariably wishes to do" (*Critique of Practical Reason*, Academy 37). Gauthier assumes that for Kant – and for any reasonable position, including his own – a practical law (as opposed to a law of nature) is *prescriptive*. It is, of course, true that according to Kant the Moral Law appears to us finite rational beings as an imperative; but this does not imply that the Moral Law (surely a practical law, if there is one) cannot be understood as the law of the *actual* operation of the pure will (as long as the pure will is in no way obstructed by anything different from it).

[130] Immanuel Kant, *Grounding for the Metaphysics of Moral*, tr. James W. Ellington (Indianapolis: Hackett, 1981), Academy 417.

[131] Kant, *Grounding*, Academy 418.

> Kant does not discern the parallel between the activities of the will and understanding and so he does not recognize that, just as the pure concepts of understanding prescribe a law a priori to appearances, so the pure concepts of the will prescribe a law a priori to our choices.[132]

Gauthier maintains that being subject to a practical law somehow based on desire is compatible with the autonomy of a rational agent. An agent that pursues happiness, an overall satisfaction of his desires, is not directly determined by these desires as they come and go (in "mere animal responsiveness to immediate need"[133]), but acts on the basis of a self-given, comprehensive framework of choice. Such an agent is determined not by his desires or inclinations as they happen to pull him in various directions, but by his will or practical reason. So, Gauthier concludes that "Kant's emphasis on and concern with autonomy is retained by our reinterpretation."[134]

We should add that for Gauthier the practical law based on happiness is not a principle prescribing the direct maximization of desire satisfaction. In his book *Morals by Agreement*, he presents a mode of maximization involving constraints on the individual pursuit of happiness ('constrained maximization') and argues that without such constraints individual attempts at maximization will prove self-defeating.[135] The details of his argument need not concern us here.

[132] Gauthier, *Moral Dealing*, 116-117.

[133] Gauthier, *Moral Dealing*, 126.

[134] Gauthier, *Moral Dealing*, 117.

[135] Gauthier, *Morals by Agreement*, ch. VI.

For our purposes it is sufficient to note that Gauthier defends the rationality of constraints on maximization by an argument from Rational Choice Theory intended to show that *constrained* rather than *direct* maximization maximizes expected utility.

4.2

In the arguments sketched in Section 4.1 Gauthier seems to follow an interpretive tradition in which Kant's distinction between heteronomy and autonomy is read as the distinction between naturally caused action and action determined by the will (i.e., practical reason). According to this interpretation, Kant's "general principle of self-love, or one's happiness"[136] states the natural necessity that all our non-moral actions have our own happiness as their end. To this is often added that Kant's notion of happiness (as a naturally given end) is essentially hedonistic.[137]

Recently, however, this reading of Kant has come under attack, especially in the work of Thomas Hill, Andrews Reath, and Henry Allison.[138] These interpreters all point out that in Kant's theory of motivation nothing can count as an action unless it follows from a *maxim*, i.e., from a general rule of practical principle the agent freely adopts for himself. According to Kant,

[136] Kant, *Critique of Practical Reason*, Academy 25.

[137] See, for instance, Lewis White Beck, *A Commentary on Kant's Critique of Practical Reason* (Chicago: The University of Chicago Press, 1960), ch. VII.

[138] Thomas Hill Jr., *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca: Cornell University Press, 1992), chs. 5, 6, and 7; Andrew Reath, "Hedonism, Heteronomy, and Kant's Principle of Happiness," *Pacific Philosophical Quarterly* 70 (1989), 42-72; Allison, *Kant's Theory of Freedom*. See also Gerold Prauss, *Kant über Freiheit als Autonomie* (Frankfurt/Main: Vittorio Klostermann, 1983).

no naturally given incentive (like, for example, anticipated pleasure) can determine a choice and action unless it has been incorporated as an end into a maxim.[139] This means that no incentive issues in an action if the agent has not made it his principle to act on the grounds of having this incentive. Thus, an agent's actions from self-love proceed from his adopted maxim of making his own happiness the determining ground of his choices. Contrary to Gauthier's claim, self-interested actions are, for Kant, not the result of natural necessity, but the expression of a freely chosen practical principle according to which the promotion of one's happiness is the ultimate reason for action.

The distinction between autonomy and heteronomy can now be seen not as the distinction between free moral action and naturally caused non-moral action, but as a distinction between two types of determinations of the will. An autonomous will determines itself by choosing its maxims independently of all naturally given desires and inclinations. In other words, for the autonomous will the presence of desires and inclinations does not provide the ultimate reason for action. This does not mean that, according to Kant, the content of inclinations cannot contribute to the specification of the ends of an autonomous will. But for the autonomous will the presence of an inclination never determines its being incorporated into a maxim as an end. The determination of an autonomous will is self-determination and thus proceeds – in some sense – from its nature as a will. A heteronomous will, on the other hand, chooses its maxims for the sake of a (possibly comprehensive) satisfaction of given desires and inclinations. In this way the will that determines itself according to the principle of self-love is heteronomous since its determination ultimately depends on influences external to itself, i.e., on the contingent presence of desires and inclinations.

[139] This is Allison's central interpretive thesis in his *Kant's Theory of Freedom*.

Given this interpretation of the autonomy/heteronomy distinction, the reason for Kant's rejection of the principle of happiness as a practical law cannot be – as Gauthier claims – that happiness is pursued not by choice, but by natural necessity. It should also be clear that in Gauthier's naturalized Kantianism 'autonomy' is not used in Kant's sense of the term. Rather than the will's capacity of determining itself independently of any given sensible desires and inclinations, 'autonomy,' for Gauthier, means only the capacity of 'standing back' from the given manifold of desire and of constructing an end of happiness out of this unstructured given multiplicity. But we may ask why no principles prescribing happiness can function as practical laws in the way intimated by Gauthier. Why *couldn't* the principles of Rational Choice Theory that unify the manifold of desire (and thereby make possible the pursuit of the end of happiness) be the a priori practical laws that are "objective, i.e., valid for the will of every rational being"?[140]

4.3

If the concepts and principles of Rational Choice Theory are understood as analogous to the concepts and principles of pure understanding, the following question presents itself immediately: what justifies the (practical) validity of these a priori concepts and principles of the will? Or, in Kant's terms: what is their deduction?

'Deduction' could here refer to different types of argument. First, in a deduction of Rational Choice Theory we might try to establish that we do have a will, i.e., a capacity of self-

[140] Kant, *Critique of Practical Reason*, Academy 19.

determination, and then argue to the validity of the principles of Rational Choice Theory (perhaps by means of a 'reciprocity thesis' like "having a will and standing under the precepts of Rational Choice Theory imply each other").[141] Second, in a metaphysically less ambitious deduction we would not argue for the existence of a will capable of self-determination, but would attempt to show that *if* we conceive ourselves as agents, we must unify the manifold of desire according to Rational Choice Theory. Such a deduction would result in a conditional conclusion: our very concept of an agent with a unified set of desires presupposes the concepts and principles of Rational Choice Theory.[142]

Here I will only consider the second, more modest version of a deduction. Although Gauthier does not explicitly address the question of how the 'pure concepts of the will' are to be deduced, he seems to have a deduction of the second type in mind.[143] This kind of deduction requires a transcendental argument to the conclusion that the principles of Rational Choice Theory are necessary condition of unified agency. In this Section I shall argue that this requirement cannot be met.

In our case the starting point of the transcendental argument for the validity of the concepts and principles of Rational Choice Theory must be the conception of a (self-determined)

[141] Henry Allison has given the name 'reciprocity thesis' to the claim (which he attributes to Kant) that having a free will and being under the moral law imply each other. See his *Kant's Theory of Freedom*, chs. 11 and 12.

[142] Note that this kind of deduction would be the practical analogue to a Strawsonian interpretation or reconstruction of the deduction of the pure concepts of the understanding. (See P.F. Strawson, *The Bounds of Sense* (London: Methuen, 1966). Such a reconstruction makes Kant's results in the Transcendental Deduction analytic, i.e., reduces them to conclusions of something like 'conceptual analysis.')

[143] See Gauthier, *Moral Dealing*, 115-116.

agent with unified desires. What is implied in this conception? Gauthier hints at the relevant considerations in the following passage:

> Such a conception requires grasping our needs as united into a single whole – as the needs of one person. Only so conceived do they give rise to the thought of happiness as a single object of desire. An animal has desires corresponding to its needs, but it is incapable of thinking of each as "I desire." Neither its intuitions nor its desires constitute a single experience because it lacks the rational capacity to unite what is given separately.[144]

In other words, for me as an agent with unified desires it is necessarily true that I can become conscious of all of my desires as belonging to my identical self. This constitutes a kind of original transcendental unity of consciousness (which we may call the 'original practical unity of consciousness'), a unity that is analogous to Kant's original transcendental unity of apperception. As the latter functions as the fundamental premise in the deduction of the pure concepts of the understanding, the former may be seen as the starting point for an attempt at a deduction of pure concepts of the will. If these pure concept are those of Rational Choice Theory, their deduction must show them to be necessary conditions for the original practical unity of consciousness through which the "I desire" is able to accompany whatever is in my motivational set.

For such a deduction the following outline of an argument might suggest a promising route. The original practical unity of consciousness presupposes the agent's capacity of coming up with choices on the basis of the manifold of desire or, to put it in the language of Rational

[144] Gauthier, *Moral Dealing*, 115.

Choice Theory, of constructing a choice set for the available alternatives over which he has preferences. If the agent did not have this capacity, he would be determined by the influence of desires or inclinations as they happen to be present. If he were so determined, his choices would issue from a ‘self’ that would be – to paraphrase Kant – as “many-colored and diverse as it has desires of which it is conscious.”[145] Rational Choice Theory provides the rules for constructing choice sets and thereby makes the practical unity of consciousness possible. These rules are given by ordinal utility theory for the case of choice under certainty and by expected utility theory for the cases of choice under risk and uncertainty.

The assumptions and inferences of a deduction along these lines obviously need further clarification and defense. Here I will discuss only one problem in the argument. In the sketched deduction it is assumed that there is no ambiguity in the concepts of rational choice and (expected) utility. In other words, it is taken for granted that Rational Choice Theory provides uniquely determined practical principles for the solution of problems of choice. Without this assumption the deduction could not establish the validity of any particular principles of choice that determine a choice set. It could only show that some such principles are necessary.[146]

Rational Choice Theory is essentially utility theory and prescribes the maximization of (expected) utility. But given the structure of utility theory, this prescription is much less determinate than it might appear. A utility theory in Rational Choice Theory has as its core a representation theorem which states that if an agent satisfies a certain set of axioms about his

[145] Compare Kant, *Critique of Pure Reason*, B 134.

[146] One might compare this to Kant’s transcendental argument for the Second Analogy. This argument can only establish universal causal connection. It does not specify the particular form of this connection. For instance, the argument by itself is not mean to establish Newtonian principles.

preferences, there exists an expected utility function representing these preferences. Therefore, the notion of utility varies with each different axiomatization utility, and different such axiomatizations lead to incommensurable notions of utility. If we are faced with different utility theories, a transcendental argument like the one outlined above cannot establish which of these theories we ought to adopt. Now, there do exist competing utility theories equally capable of unifying the manifold of desire by determining choice sets. If these theories are taken to explicate the notion of choosing for the sake of happiness (interpreted as utility, i.e., the satisfaction of our desires or preferences), they present us with different conceptions of happiness. These different conceptions of happiness are tied to different practical principles for which it now seems impossible to give a transcendental justification that takes as its starting point the practical unity of consciousness.

This general argument is best illustrated by a brief comparison of two distinct theories of expected utility, one based on the work of John von Neumann, Oskar Morgenstern, and Leonard Savage (which I will refer to as the 'Standard Theory'), the other developed by Richard Jeffrey and Ethan Bolker ('Jeffrey/Bolker Theory').[147] Note that I discuss these particular theories only to clarify a general problem about using utility theory in the explication of the notion of a

[147] For an exposition of the Standard Theory, see John C. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge: Cambridge University Press, 1977), ch. 3. For the Jeffrey/Bolker Theory, see Richard Jeffrey, *The Logic of Decision* (Chicago: The University of Chicago Press, 1983) and Ethan Bolker, “A Simultaneous Axiomatization of Utility and Subjective Probability,” *Philosophy of Science* 43 (1967), 333-340; also John Broome, “Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism,” *Review of Economic Studies* 57 (1990), 477-502.

practical law based on happiness. I could make the same point by comparing, say, ‘evidentiary’ and ‘causal’ decision theory.[148]

The two theories differ mainly in their representation of uncertain prospects. In the Standard Theory an uncertain prospect P is modeled as a contingency mixture or gamble:

$$P = (A_1/e_1; A_2/e_2; \ldots; A_k/e_k)$$

Under the uncertain prospect P, outcome A_1 will occur if event (or ‘state of nature’) e_1 occurs; A_2 if e_2, etc. The events $e_1, \ldots, e_k$ are chosen so that one of them and only one of them will occur. In other terms, $e_1, \ldots, e_k$ stand for a set of k mutually exclusive and exhaustive possibilities. For example, in the gamble of my betting that Grasshoppers Zurich will win the Europa League there are two states of nature (“Grasshoppers win” and “Grasshoppers do not win”) and two outcomes (“I win $10” and “I lose $10”). In the Standard Theory the preference relation is defined for such gambles. It is assumed that given a set of outcomes and a set of states of nature, an agent has a preference relation for all gambles that can be constructed by using the outcomes and states of nature of these sets. If the agent satisfies Savage’s axioms, the utility of the gamble $G = (A/e_1; B/e_2)$ can be represented as $U(G) = pU(A)+(1-p)U(B)$, where p is the agent’s subjective probability for the event e_1, 1-p that of the complementary event e_2.

Contingency mixtures or gambles play no role in the Jeffrey/Bolker Theory. There, prospects are modeled as propositions to which the propositional calculus can be applied. For

[148] On the distinction between evidentiary and causal decision theories see, for instance, the papers in Peter Gärdenfors and Nils-Eric Sahlin (eds.), *Decision, Probability, and Utility* (Cambridge: Cambridge University Press, 1988), Part V.

example, the uncertain prospect modeled in the Standard Theory as the Gamble G could be taken in the Jeffrey/Bolker approach as the proposition A, which is equivalent to (A&E)v(A&~E), where A might stand for the proposition "I bet that Grasshoppers Zurich will win the Europa League" and E for "Grasshoppers Zurich win the Europa League." If the agent satisfies the axioms of the Jeffery/Bolker Theory, the utility (or 'desirability,' as Jeffrey calls it) of the proposition *A* can be represented as follows:

$$D(A) = [D(A\&E)\text{prob}(A\&E)/\text{prob}(A)] + [D(A\&{\sim}E)\text{prob}(A\&{\sim}E)/\text{prob}(A)]$$

Here prob(A) is the agent's subjective probability for the proposition A, prob(A&E) that for A&E, etc. Note that in the Jeffrey/Bolker Theory the expected utility of (A&E)v(A&~E) depends on the conditional probabilities of E given A (i.e., prob(A&E)/prob(A)) and ~E given A (i.e., prob(A&~E)/prob(A)). Thus if A is interpreted as a proposition about my action, the probabilities by which D(A&E) and D(A&~E) are multiplied in order to determine the expected utility D(A) depend on the probability of my action. In this respect the setup of the Standard Theory is very different. There, probabilities are not defined for actions (i.e., gambles), but only for the states of the world (or 'states of nature'). So, the probabilities of the states of the world that determine the expected utility U(G) of my action or gamble G are entirely independent of my choice.

Both of these sketched theories of expected utility (or 'desirability') unify the manifold of desire by determining choice sets for uncertain prospects. According to both, the rational choice is a maximizing choice, and both contain a principle of maximizing expected utility. Yet in the Standard Theory the principle does not mean what it means in the Jeffrey/Bolker Theory. In important types of decision situations the two theories prescribe different choices of action as the utility maximizing choice. For instance, in decision situations with the structure of a Prisoner's Dilemma or a Newcomb problem, the Standard Theory always prescribes choosing non-cooperation and taking the so-called 'two-boxes' option, whereas the Jeffrey/Bolker Theory may recommend cooperation and taking the 'one-box' option. This important difference is due exactly to the different ways, outlined above, in which the two theories model uncertain prospects.

Consider, for instance, the Prisoner's Dilemma in which you play against your twin, i.e., someone very much like you of whom you believe that in all likelihood he will act just as yourself. You find yourself in a decision situation that can be represented as follows:

	Your twin confesses	Your twin does not confess
You confess	x	y
You do not confess	z	w

Your preferences over the outcomes x, y, z, w, are yPwPxPz (where P stand for the strict preference relation).[149]

According to the Standard Theory, you should always confess in situation with this structure. You face a choice between two uncertain prospects that can be modeled as gambles: confession is the gamble (x/your twin confesses; y/your twin does not confess); non-confession

[149] Little depends on the artificial character of this particular decision situation. I have chosen it only for the sake of simplicity. For more realistic examples of situations with this Prisoner's Dilemma structure or the structure of a Newcomb problem, see Broome, 487-488,

the gamble (z/your twin confesses; w/your twin does not confess). No matter what the ‘states of nature’ (the events of your twin’s confession or non-confession, which – according to the Standard Theory – are independent of your choice) turn out to be, you do better by confessing. Confessing is the *dominant* choice. Since the axioms of the Standard Theory imply a dominance principle, you cannot conform to the theory unless you confess.

The Jeffrey/Bolker Theory, on the other hand, recommends in most cases of this Prisoner’s Dilemma that you do not confess. Let A stand for the proposition “You confess”; B for “Your twin confesses”; ~A for “You do not confess”; and ~B for “Your twin does not confess.” Your preferences are (A&~B)P(~A&~B)P(A&B)P(~A&B). You face a choice between uncertain prospects expressed in propositions A and ~A. A and ~A are equivalent, respectively, to (A&B)v(A&~B) and (~A&B)v(~A&~B). The expected utilities (or ‘desirabilities’) of your confession and non-confession are:

$$D(A) = [D(A\&B)\mathrm{prob}(A\&B)/\mathrm{prob}(A)] + [D(A\&{\sim}B)\mathrm{prob}(A\&{\sim}B)/\mathrm{prob}(A)]$$

$$D({\sim}A) = [D({\sim}A\&B)\mathrm{prob}({\sim}A\&B)/\mathrm{prob}({\sim}A)] + [D({\sim}A\&{\sim}B)\mathrm{prob}({\sim}A\&{\sim}B)/\mathrm{prob}({\sim}A)]$$

Since you know that you play your twin, the conditional probabilities prob(A&B)/prob(A) and prob(~A&~B)/prob(~A) are very high. Therefore, D(A) is near D(A&B) and D(~A) near D(~A&~B). Since (A&B)P(~A&~B), you ought not to confess (unless your preferences are such that the difference between D(A&B) and D(~A&~B) is very small, and that between D(~A&B) and D(A&~B) very large).[150]

Thus, in a situation with such a structure the principle of maximizing expected utility is indeterminate as long as it is not specified which utility theory we are to apply. Both confession and non-confession can be represented as maximizing expected utility. These representations are, of course, relative to different utility theories, but – and this is the crucial point – there is no notion of maximal expected utility that is not relative to a particular utility theory. Different utility theories do not specify different ways of achieving a goal (maximal utility or happiness) whose content is determined independently of these theories. Once we identify happiness and maximal expected utility, we tie our notion of happiness to some utility theory, and can no longer assess rival utility theories by appealing to this very notion. (We may add here that in *Morals by Agreement* Gauthier argues for the superiority of *constrained* over *straightforward* maximization of utility: constrained maximizers will be better off than straightforward maximizers because they will be able to reap the benefits of cooperation in Prisoner's Dilemma situations.[151] Gauthier's argument depends, however, on the Standard Theory: constrained maximizers are supposed to get more *utility* (the notion of which Gauthier takes from the familiar von Neumann/Morgenstern framework) than straightforward maximizers. But by cooperating in

[150] We must note that Richard Jeffrey recommends the 'ratifiable' choice in the Prisoner's Dilemma, i.e., the choice of confession. (On ratifiability, see Jeffrey, 15-25). However, as an addition to the Jeffrey/Bolker Theory ratifiability seems entirely ad hoc and ill-connected with the axiomatic development of the theory. If the Jeffrey/Bolker Theory is meant to be a general theory of rationality, the rationality of confession in the Prisoner's Dilemma ought to be derivable form the axioms of the theory and not from some ad hoc principle added in order to accommodate a prior conviction about the best course of action in a particular type of situation.

[151] Gauthier, *Morals by Agreement*, ch. VI.

Prisoner's Dilemmas, constrained maximizers violate the axioms that define the notion of utility Gauthier accepts. Now, it is simply impossible to become a better utility maximizer by violating the very axioms of one's own conception of utility.[152])

One might object that in our comparison of the two utility theories we have not applied them to the same decision situation and that this explains the appearance of their incompatibility. It is easy to see that our initial description of the Twin Prisoner's Dilemma corresponds neither to the mode of representing uncertain prospects of the Standard Theory nor to that of the Jeffrey/Bolker Theory. (In our first presentation of the situation we let preferences range over outcomes and not over gambles or propositions.) So, it could be claimed that since a Prisoner's Dilemma is *defined* as a situation in which it is always rational to confess, the situation to which we have applied the Jeffrey/Bolker Theory is simply not a Prisoner's Dilemma.

We can grant the terminological point about what we should call a 'Prisoner's Dilemma,' but this leaves open the question of whether or not we *should* represent a given decision situation as a Prisoner's Dilemma. After all, decision situations do not come with ready-made recipes for their representation. The adoption of a particular utility theory commits us to a particular way of representing decision situations that is incompatible with that of a different theory. As we have seen, different representations of uncertain prospects in different utility theories lead to different prescriptions of courses of action.

A particular utility theory can be advocated and defended against its competitors by a variety of reasons,[153] but not simply for the reason that it unifies the manifold of desire and

[152] For more on this point see my "Expected Utility and Constrained Maximization: Problems of Compatibility," *Erkenntnis* 41 (1994), 37-48.

makes the practical unity of consciousness possible. More than one worked-out utility theory may perform this task. So, the proposed transcendental argument which was our starting point in this Section cannot specify the precise nature of the utility theory we need to adopt, i.e., the rules according to which we have to unify the manifold of desire. If these rules remain unarticulated, Gauthier's original claim, that through the principles of utility theory we obtain determinate practical laws based on happiness, seems no longer defensible. As we have seen, a practical law simply prescribing the pursuit of happiness (interpreted as the maximization of expected utility) could not even tell us what to do in a Prisoner's Dilemma.[154] Thus, the attempt to justify principles of Rational Choice Theory in the same way in which Kant tries to justify the principles of pure understanding appears rather unpromising. (Here we have not considered Rational Choice Theory for strategic decision situations, i.e., Game Theory. But with respect to Game Theory we could develop an argument very similar to that of this Section. We could argue that the prescription to choose equilibrium strategies – the prescription central to Game Theory – is indeterminate given the multiplicity (if not profusion) of well-defined and axiomatically

[153] One might favor the Standard Theory over the Jeffrey/Bolker Theory because one might be uneasy about the fact that in the latter the utility of an action depends on the probability of the *action* itself.

[154] With his own version of Rational Choice Theory, i.e., the theory of *constrained maximization*, Gauthier claims to have 'solved' the Prisoner's Dilemma (since constrained maximizers will cooperate). But this pretended solution does not imply that followers of the Standard Theory (who do not cooperate) fail to maximize utility. (Moreover, it is not even clear what it is that constrained maximizers maximize: after all, in cooperating in Prisoner's Dilemmas they violate the axioms of the Standard Theory.)

characterized equilibrium concepts (as well as the possible multiplicity of equilibria, once we accept a particular equilibrium concept).[155])

4.4

In his reinterpretation of Kant, Gauthier tries to retain a practical role for reason, i.e., the role of a mediator between the manifold of desire and choice.[156] Gauthier also assumes that thereby he preserves Kant's concern with autonomy. As we have pointed out in Section 4.2, Kant's own notion of autonomy is different from and stronger than the one Gauthier ascribes to him. I shall argue now that once we accept a practical role for reason, even if only for the task of filling the "gap between desire and action,"[157] we are led towards a Kantian notion of autonomy.

In Section 4.3 we have shown that an interest in the maximal satisfaction of our desires cannot provide a determinate principle for the unifying activity of the will or practical reason (as understood by Gauthier). Reflection on the account of motivation (the account of reasons for action) we must accept if we are to allow for a practical role of reason, shows that this should not strike us as a surprising result.

Gauthier's view on the task of practical reason commits him to a theory of motivation that takes a middle position – an unstable one, as we shall see – between (so-called) Humean and

[155] For a good overview of the major equilibrium concepts, see Roger Myerson, *Game Theory* (Cambridge/MA: Harvard University Press, 1991). (On the multiplicity of equilibria, see especially chs. 4 and 5.)

[156] We need not to decide here whether this formulation is adequate to *Kant's* understanding of the practicality of reason.

[157] Gauthier, *Moral Dealing*, 126

Kantian approaches. On the one hand, reasons for action arise originally from given desires and direct us to the maximal satisfaction of them. But, on the other hand, these given desires must first be judged by practical reason so that they can become components of a coherent whole, i.e., a conception of happiness. This means, however, that no given desire is a reason for action unless it is taken up by practical reason into an ordered system.[158] On this view, we are therefore not motivated by desires as they happen to make themselves felt, but by a kind of judgment of practical reason about these desires. If this is the case, it is no longer clear why a maximal satisfaction of desire should be the principle of practical reason. One common argument in support of maximal desire satisfaction, i.e., an argument based on the claim that only given desires have motivational force, obviously cannot be used here. (One might now object that the principle of maximal desire satisfaction must be understood as requiring the satisfaction not of given, but of critically *judged* desires. But in this form it could not be the highest principle of practical reason since it would be conditioned by prior judgments on the rational acceptability of given desires, judgments that could not be themselves based on considerations of happiness.[159])

In unifying the manifold of desire, practical reason is to judge which desires ought to be satisfied. In order to attain such unification, reason's judgment must be directed by a principle. The question then arises: what is this principle, and how is it to be justified? Whatever the principle of practical reason might be, it cannot be based on given desires since – by hypothesis –

[158] In Thomas Nagel's terms, only *motivated* desires are reasons. See his *The Possibility of Altruism* (Oxford: Clarendon Press, 1970). (In a more Kantian comparison, we could say that just as given sensations do not by themselves constitute knowledge, given desires do not by themselves constitute reasons for action.)

[159] Stephen L. Darwall argues that the demand of maximal satisfaction of critically *judged* desires is compatible with the demand of Kant's Categorical Imperative. See "Kantian Practical Reason Defended," *Ethics* 96 (1985), 89-99.

no such desire constitutes a reason. If practical reason is to decide which of our given desires or preferences are reasons for action, its decisions cannot depend on any of those desires whose reasonableness is at issue. The judgment of practical reason cannot be based on given second-order or higher-order desires, either: such higher-order desires belong to the manifold of desire and may very well be in conflict with each other. Moreover, it is an open question whether in a conflict between a second-order and an opposed first-order desire the higher-order desire should prevail. If there is to be a balancing of given first-order and higher-order desires (or a decision between them), this is precisely the task of practical reason.

If the unification of the manifold of desire by practical reason is not determined by given desires, it is to be expected that an interest in maximal desire satisfaction should not give us a definite practical principle. We have seen that the *very meaning* of happiness as maximal desire satisfaction, i.e., as utility maximization, is relative to particular and incompatible theories of utility.[160]

It might clarify matters to distinguish my point here from the claim that happiness cannot give rise to a practical law because of the subjective variability of desires or preferences.[161] The point is not that, because people differ in their desires, there could be no practical law with a universal applicability. Even though desires vary from person to person, the demand that they be maximally satisfied – whatever they happen to be – still seems unequivocal. The problem is

[160] As Sergio Tenenbaum pointed out to me, a hard-nosed subjectivist about reasons for action could attempt a relativism extending to frameworks of maximization. But this would undermine a basis of support of subjectivism: the seemingly unambiguous notion of acting for the satisfaction of one’s desires.

[161] For a classic version of this point (made against utilitarianism), see F.H. Bradley’s *Ethical Studies* (Oxford: Clarendon Press, 1927), Essay III.

rather that in order to maximize one's desire satisfaction or utility one has to adopt a framework of maximization, i.e., a particular utility theory. As explained in Section 4.3, there are situations in which two agents with exactly the same given desires might maximize their desire satisfaction by choosing different courses of action (which will be determined by their adoption of incompatible theories of expected utility). Under the view of motivation we are considering here, nothing in the manifold of desire compels practical reason to adopt one of these theories rather than another. More importantly, nothing in the manifold of desire is a reason to adopt *any* such theory of maximization at all.[162]

My argument here can be summarized as follows: if the task assigned to practical reason is that of mediating between the manifold of desire and a conception of happiness, practical reason needs to perform this task according to a determinate practical principle. Since a concern with happiness does not give rise to a unique such principle, there must be a justification for adopting one particular principle of happiness rather than another. But whatever this justification will turn out to be, it cannot come from given desires. Thus, it may now appear plausible that it must, in some way, depend on the nature of practical reason itself. But if the justification of a principle of practical reason must be of such a form, the will that adopts this principle must be autonomous in *Kant's* sense of 'autonomy.' Here we need not – and, of course, cannot – specify what the principle of an autonomous will is; but given the development of the notion of an autonomous will in Kant and his successors, it seems unlikely that it should prove to be a principle of happiness.

[162] Similarly, it is the understanding itself that with its spontaneous legislation unifies the manifold of intuition. This legislation is in no way determined by given sensations.

4.5

We can draw the conclusion that once we interpret utility theory as a theory of reasons for action, a Kantian conception of motivation is no longer optional. According to utility theory as a theory of reasons, it is necessary that a rational agent be able to form a coherent whole of preferences out of his given desires. But we can ascribe this ability to the agent only if we assume him to a capacity of judging his desires independently of the given manifold of desire, i.e., if we assume him to have a faculty of practical reason (*pure* practical reason – as Kant would put it). Thus, we can take it – although somewhat subversively – as the lesson of Gauthier's reinterpretation of Kant that an account of practical rationality as utility maximizing presupposes a fully Kantian conception of autonomous agency.[163]

At any rate, it appears that moral philosophers cannot use utility theory in subjectivist accounts of value. Subjectivists about value agree that – in some sense of 'objective' – there are not objective values and argue that values or reasons for action must be tied to our given desires or other given motivational states. For instance, Gautier formulates the main claim of subjectivism as follows:

> Value is then not an inherent characteristic of things or states of affairs, not something existing as part of the ontological furniture of the universe in a manner quite independent of persons and their activities. Rather, value is created or determined through preference.

[163] For an argument in a similar direction, see Stephen L. Darwall, *Impartial Reason* (Ithaca: Cornell University Press, 1983), ch. 6.

> Values are products of our affections. To conceive of value as dependent on affective relationships is to conceive of value as *subjective*.[164]

Gauthier then points to the main difference between a subjective and an objective conception of value:

> To conceive of value as objective is to conceive of it as existing independently of the affections of sentient beings, and as providing a norm or standard to govern their affections. The subjectivist denies the existence of such a norm.[165]

Similarly, J.L. Mackie claims that, according to subjectivism, there are no practical principles which are "unconditional in the sense of not being contingent upon any present desire of the agent."[166]

But if only given, present desires constitute reasons for action, utility theory cannot furnish any practical principles. As we have argued, insofar as utility theories do provide norms or standards to govern our desires, they cannot be derived from the given manifold of desire. Thus, a subjectivist cannot accept the maximization of utility as the norm or standard of choice.

The only way a utility theory could play a role in a subjectivist account of value and reasons for action would be as a part of a highly idealized theory of reasoning processes of

[164] Gauthier, *Morals by Agreement*, 47.

[165] Gauthier, *Morals by Agreement*, 47. (We might wonder here whether Gauthier would accept that there is no notion of value for non-sentient beings, beings without (presumably sensible) affections.)

[166] Mackie, *Ethics – Inventing Right and Wrong*, 29.

agents faced with a manifold of desire, i.e., a theory of how the agent's unstructured affections and desires are transformed into coherent – or, at least, largely coherent – preferences governing his choice. A utility theory would thus contribute to the explanation of determinate preferences and choices. In this way it would meet the requirement – proposed, for example, by Gilbert Harman – that a subjectivist theory of reasoning be (part of) an empirical theory: "in any event, the theory of reasons is an empirical theory – that is the important point."[167]

As part of an empirical theory of reasoning, utility theories had better be well confirmed. But, as is widely acknowledged, this appears not to be the case. According to a great many experimental tests, agents seem systematically to violate the axioms of these theories.[168] Even if we granted the empirical adequacy of a utility theory, the theory could not provide any practical principles (except as descriptive causal principles). It is, of course, open to the subjectivist to maintain that, strictly speaking, there are no practical principles, but only causal principles of actual reasoning processes. But this is clearly not an option for Gauthier if, as we have seen, he wants to defend the principles of utility theory as a priori practical principles that are analogous to the a priori principles of the (theoretical) understanding. When Gauthier claims that there are no objective norms or standards, he does not imply that there are no norms or standards (and corresponding practical principles) whatsoever. Moreover, given a causal view of reasoning, a subjectivist can hardly claim – a Mackie does – that "morality is not to be discovered but to be made: we have to decide what moral views to adopt, what moral stands to take."[169] To adopt a

[167] Harman, *The Nature of Morality*,131.

[168] The literature on this topic is vast. For some important papers and a bibliography, see Gärdenfors and Sahlin, *Decision, Probability, and Utility*.

[169] Mackie, *Ethics – Inventing Right and Wrong*, 106.

moral stand must at least mean to decide which of one’s desires should count as reasons for one’s action. We cannot see ourselves as making such a decision without attributing to ourselves a capacity to act on practical principles. Otherwise, there seems to be no way in which the adoption of a moral (or any practical stand) can be taken as reasonable.

BIBLIOGRAPHY

Allison, Henry. *Kant's Theory of Freedom.* Cambridge: Cambridge University Press, 1990.

Árdal, Páll S. *Passion and Value in Hume's Treatise.* Edinburgh: Edinburgh University Press, 1989.

Augustine. *Confessions*, trans. William Watts. Cambridge, MA: Loeb Classical Library, 1912.

Baier, Annette. *A Progress of Sentiments*. Cambridge, MA: Harvard University Press, 1991.

Beck, Lewis White. *A Commentary on Kant's Critique of Practical Reason*. Chicago: The University of Chicago Press, 1960.

Berkeley, George. *A Treatise Concerning the Principles of Human Knowledge*, in *The Works of George Berkeley*, ed. A.A. Luce and T.E. Jessop. London: Thomas Nelson, 1949.

Bolker, Ethan. "A Simultaneous Axiomatization of Utility and Subjective Probability." *Philosophy of Science* 43 (1967), 333-340.

Bolton, Martha Brandt. "The Taxonomy of Ideas in Locke's *Essay,*" in *The Cambridge Companion to Locke's "Essay concerning Human Understanding,"* ed. Lex Newman. Cambridge: Cambridge University Press, 2007.

Bradley, F.H. *Ethical Studies*. Oxford: Clarendon Press, 1927.

Brandt, Richard B. *A Theory of the Right and the Good*. Oxford: Clarendon Press, 1979.

Cudworth, Ralph. *The True Intellectual System of the Universe*. London: Royston, 1678.

Broome, John. "Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism." *Review of Economic Studies* 57 (1990), 477-502.

Darwall, Stephen L. *Impartial Reason*. Ithaca: Cornell University Press, 1983.

Darwall, Stephen L. "Kantian Practical Reason Defended." *Ethics* 96 (1985), 89-99.

Driver, Julia. *Uneasy Virtue*. Cambridge: Cambridge University Press, 2001.

Franklin, Benjamin. *The Autobiography*. New York: First Vintage Books/The Library of America, 1990.

Gärdenfors, Peter and Sahlin, Nils-Peter (eds.). *Decision, Probability, and Utility*. Cambridge: Cambridge University Press, 1988.

Gauthier, David. *Morals by Agreement*. Oxford: Clarendon Press, 1986.

Gauthier, David, *Moral Dealing*. Ithaca: Cornell University Press, 1990.

Harman, Gilbert. *The Nature of Morality*. Oxford: Oxford University Press, 1977.

Harsányi, John C. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press, 1977.

Henrich, Dieter. "Der Begriff der sittlichen Einsicht und Kants Lehre vom Faktum der Vernunft," in *Kant: Zur Deutung seiner Theorie von Erkennen und Handeln*, ed. Gerold Prauss. Köln: Kiepenheuer & Witsch, 1973.

Hill, Thomas. *Dignity and Practical Reason in Kant's Moral Theory*. Ithaca: Cornell University Press, 1992.

Hobbes, Thomas. *Leviathan*, ed. C.B. Macpherson. Harmondsworth: Penguin, 1968.

Hume, David. *A Treatise of Human Nature*, 2nd ed. L.A. Selby Bigge and P.H. Nidditch. Oxford: Clarendon Press, 1978.

Hume David. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford: Clarendon Press, 1975.

Hume, David. *Essays Moral, Political, and Literary*, ed. Eugene Miller. Indianapolis: Liberty Classics, 1985.

Hume, David. *Dialogues Concerning Natural Religion*, ed. Norman Kemp Smith. Indianapolis: Bobbs-Merrill, 1947.

Jeffrey, Richard. *The Logic of Decision*. Chicago: The University of Chicago Press, 1977.

Kant, Immanuel. *Critique of Pure Reason*, trans. Norman Kemp Smith. London: Macmillan, 1929.

Kant, Immanuel. *Critique of Practical Reason*, trans. Lewis White Beck. Indianapolis: Bobbs-Merrill, 1956.

Kant, Immanuel. *Grounding for the Metaphysics of Morals*, trans. James W. Ellington. Indianapolis: Hackett, 1981,

Kemp Smith, Norman. *The Philosophy of David Hume*. London: Macmillan, 1941.

King, Lord Peter. *The Life of John Locke*. London: Colburn and Bentley, 1830.

Krook, Dorothea. *Three Traditions of Moral Thought*. Cambridge: Cambridge University Press, 1959.

Leibniz, Georg Wilhelm. *New Essays on Human Understanding*, trans. and ed. Peter Remnant and Jonathan Bennett. Cambridge: Cambridge University Press, 1982.

Lennon, Thomas. "Locke and the Logic of Ideas." *History of Philosophy Quarterly* 18 (2001), 155-176.

Locke, John. *An Essay Concerning Human Understanding*, ed. Peter H. Nidditch. Oxford: Clarendon Press, 1975.

Locke, John. *Political Essays*, ed. Mark Goldie. Cambridge: Cambridge University Press, 1997.

Lottenbach, Hans. "Expected Utility and Constrained Maximization: Problems of Compatibility." *Erkenntnis* 41 (1994), 37-48.

MacIntyre, Alasdair. *Whose Justice? Whose Rationality?* Notre Dame, IN: University of Notre Dame Press, 1988.

Mackie, J.L. *Ethics – Inventing Right and Wrong*. Harmondsworth: Penguin, 1977.

Mackie, J.L. *Hume's Moral Theory*. London: Routledge & Kegan Paul, 1980.

Malebranche, Nicolas. *De la Recherche de la Vérité*, in *Oeuvres complètes de Malebranche*, ed. André Robinet, tomes I-III. Paris: J. Vrin, 1962-1964.

Myerson, Roger. *Game Theory*. Cambridge, MA: Harvard University Press, 1991.

Prauss, Gerold. *Kant über Freiheit als Autonomie.* Frankfurt/Main: Vittorio Klostermann, 1983.

Nagel, Thomas. *The Possibility of Altruism*. Oxford: Clarendon Press, 1970

Reath, Andrews. "Hedonism, Heteronomy, and Kant's Principle of Happiness." *Pacific Philosophical Quarterly* 70 (1989), 42-72.

Rousseau, Jean-Jacques. *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*, in *Oeuvres complètes*, tome III, ed. Marcel Raymond and Bernard Gagnebin. Paris: Gallimard, 1964.

Sales, François de. *Traité de l'Amour de Dieu*, in *Oeuvres*, ed. André Ravier. Paris: Gallimard, 1969.

Siebert, Donald T. *The Moral Animus of David Hume*. London and Toronto: Associated Universities Press, 1990.

Strawson, P.F. *The Bounds of Sense*. London: Methuen, 1966.

Stroud, Barry. *Hume*. London: Routledge & Kegan Paul, 1977.

Thiel, Udo. *Lockes Theorie der personalen Identität*. Bonn: Bouvier, 1983.

Williams, Bernard. "Internal and External Reasons," in *Moral Luck*. Cambridge: Cambridge University Press, 1981.

Williams, Bernard. *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press, 1985.