POZNAŃ STUDIES
IN THE PHILOSOPHY OF THE SCIENCES AND THE HUMANITIES

94

MORAL PSYCHOLOGY

Edited by
Sergio Tenenbaum

Rodopi

MORAL PSYCHOLOGY

POZNAŃ STUDIES
IN THE PHILOSOPHY OF THE SCIENCES AND THE HUMANITIES

VOLUME 94

Poznań Studies in the Philosophy of the Sciences and the Humanities
is sponsored by SWPS

Address: dr hab. Katarzyna Paprzycka, Prof. SWPS · Department of Philosophy · SWPS
ul. Chodakowska 19/31· 03-815 Warszawa · Poland · fax: ++48 22 517-9625
E-mail: PoznanStudies@swps.edu.pl · Website: http://PoznanStudies.swps.edu.pl

NEW TRENDS IN PHILOSOPHY

Katarzyna Paprzycka (editor-in-chief)
Department of Philosophy · SWPS
ul. Chodakowska 19/31 · 03-815 Warszawa · Poland
Katarzyna.Paprzycka@swps.edu.pl

New Trends in Philosophy is a new subseries of the *Poznań Studies in the Philosophy of the Sciences and the Humanities* book series. It publishes collections of papers that deal with new or underrepresented topics in philosophy.

Other volume in the series:

Vol. 92: M.P. Wolf and M.N. Lance (eds.), *The Self-Correcting Enterprise: Essays on Wilfrid Sellars*

POZNAŃ STUDIES IN THE PHILOSOPHY OF THE SCIENCES AND THE HUMANITIES, VOLUME 94
NEW TRENDS IN PHILOSOPHY

MORAL PSYCHOLOGY

Edited by

Sergio Tenenbaum

Amsterdam - New York, NY 2007

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISSN 0303-8157
ISBN-13: 978-90-420-2226-3

Printed in The Netherlands

CONTENTS

Sergio Tenenbaum

INTRODUCTION

The classic question of moral psychology concerns the nature of moral motivation and moral reasons to act. The query that ethics professors put to their first-year students – *Why be Moral?* – has not only served as the focus of perennial debate but has also led to many further questions. One wants to know what counts as a reason for action, the connection between reasons for action and motivation, whether there is any essential difference between moral and non-moral reasons, whether moral motivation has to be rational in nature or not, and even whether a sharp separation can be maintained between reasons and sentiments in the realm of practical reason. In the last few decades, the field of moral psychology has expanded to include debates that go far beyond its original motivating question. The purpose of this volume is to bring the reader scholarship from philosophers with widely divergent approaches to moral philosophy, and to serve as a sampler of the current richness of this field.

More conspicuously than in any other field in philosophy, in moral psychology the interpretation of historical texts has been inseparable from the conceptual advances made in recent decades. So it seemed that it would become such a volume to have historical essays dedicated to at least some of the historical figures whose work on moral psychology informs so much of the current debate in the field.

James Doyle's "Desire, Power and the Good in Plato's *Gorgias*" provides an interpretation and defense of Socrates' claim that the tyrants have no real power. The defense of this claim depends on Socrates' argument for an important thesis on the nature of desire, and, in particular, for the claim that what we really want is what is actually good. According to Doyle, on the Socratic view, one's real desire is for one's real good; the conception of his good determines what the rational agent desires, rather than the other way around. Meanwhile, one might think that a view that takes all desire, and *a fortiori* sensible desire, to have the

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 9-13. Amsterdam/New York, NY: Rodopi, 2007.

good as its object would be far from Kant's view. One might think that, for Kant, the aim of sensible desire is the agreeable, *rather than* the good. However, the Socratic and Kantian views might not be so sharply opposed. The relation between the agreeable and the good for Kant is examined in Stephen Engstrom's paper "Kant on the Agreeable and the Good." Engstrom argues that although the object of sensible desire is the agreeable, under certain conditions, the agreeable can be subsumed under the concept of the good. In particular, the agreeable is subsumed under the concept of the good when it is the product of practical knowledge in accordance with the idea of a practical law.

Iakovos Vasiliou's "Virtue and Argument in Aristotle's Ethics" discusses the role Aristotle sees for knowledge and argument in becoming good. Although quite a bit of emphasis has been put on the fact that Aristotle seems to think that one cannot argue someone into being virtuous, it also seems to be the case that Aristotle does not want to deny that arguments have some role to play in ethics. Vasiliou argues that the Doctrine of the Mean is not supposed to provide the agent with a way to determine what the virtuous actions are, but instead to provide an agent with what Vasiliou calls an "aiming principle." With the help of this idea of aiming principles, Vasiliou can then make a case that arguments have an important role to play in moral development, by helping agents deliberate about aiming questions.

Donald Ainslie's "Character Traits and the Humean Approach to Ethics" develops a new answer to the problem of how Hume thinks we can attribute character traits to persons, given that we *never* observe character traits. Ainslie argues that a proper answer to this question will explain various features of Hume's moral theory. For instance, in light of Ainslie's account we can see how, for Hume, there is an element of normativity in character trait ascription, and in the classification of character traits as virtuous and vicious. Although this normativity is grounded in social customs and practices and thus cannot completely depart from them, it also does not completely rule out the possibility that we take a critical stance towards it.

The theme of character traits and their importance in explaining actions is taken up in Fred Schueler's "Rationality and Character Traits." Schueler argues that traditional accounts of action explanation must stop short of their goal of explaining action. In particular, they cannot show why an agent acted for a set of reasons as opposed to another set that would justify a different action. Schueler argues that a full explanation of action requires appeal to character traits that can explain why an agent chose to act for some reasons rather than others.

David Sobel's "Practical Reasons and Mistakes of Practical Rationality" tries to defend subjectivism against a familiar objection. A subjectivist view of practical reason says, roughly, that the agent's desires or preferences determine what the agent has reasons to do. However, no subjectivist thinks that *every* desire of the agent gives her a reason (and certainly not an overriding reason) to act, or that the agent has *never* has a reason to do something that she does not currently desire. Subjectivists typically think that desires give rise to reasons only if they are well-informed desires, and that the agent might have a reason to do something on account of the fact that he would have a certain desire if he were to gain a certain piece of information. Although a subjectivist probably would want to distinguish at least between good and bad ways of processing information, many philosophers think that drawing this kind of distinction is incompatible with the subjectivist position. Sobel's paper tries to show that this objection is misguided; the subjectivist does have the resources to distinguish between good and bad information processing.

In his "Goals, Wishes, and Reasons for Action," Stephen Arkonovich also takes up the subject of subjectivism, or the Humean theory of normative reasons, from a very different perspective. Michael Smith's work has drawn attention to the often overlooked distinction between a Humean theory of *motivating* reasons (roughly, reasons that *explain* the agent's behaviour) and a Humean theory of *normative* reasons (roughly, reasons that would *justify* a certain action). Smith has famously defended what Arkonovich calls Hybrid Humeanism, the view that we should accept the Humean theory of motivating reasons, but not Humean theory of normative reasons. Although Smith's position seems to be able to incorporate the most plausible elements of Humeanism, while rejecting its most counter-intuitive aspects, Arkonovich argues that Hybrid Humeanism cannot be sustained. The implications of the Humean theory of motivation are also addressed by Philip Clark's "How Can a Reason Be Practical: A Reply to Hume." Many philosophers think that one can accept both that value judgments are objective and that they have, to use Clark's words, a "necessary grip on the will," only if we reject Hume's view that beliefs cannot motivate. For many philosophers this means that the costs of accepting these two theses about value judgments are prohibitively high given the plausibility of Hume's claim that beliefs cannot motivate on their own. Clark argues that one can keep the two theses intact without having to deny Hume's theory of motivating reasons, if we can show that practical reasoning has a constitutive goal that explains the motivational force of value judgments.

The idea that someone follows the requirements of rationality just in case they respond to reasons sounds almost like a truism. However, in "Is There a Nexus between Reasons and Rationality?", Michael Smith argues that this is false in the theoretical realm, at least if one means by "reasons," reasons that *justify* (rather than explain) the beliefs in question. Although it is true that insofar as one forms beliefs according to requirements of rationality one responds to what *seems* to one to be reasons, these might fail to be reasons even when the agent is fully rational. In the practical realm, it seems possible to hold the view that there are no reasons that justify having certain desires *tout court*, but only rational requirements that we have certain desires, give some other desires that we also have. According to Smith, there are only two ways to reject the extreme conclusion. One must either claim that there are certain beliefs such that if the agent has the belief in question than the agent also *has* certain desires (what Smith calls "the besire strategy"), or one must claim that there are certain beliefs such that if one has the belief in question then *reason requires* that the agent also have certain desires (the "rationalist strategy"). Smith argues that of the three options (the extreme view, the besire strategy, and the rationalist strategy), only the rationalist strategy is both plausible and stable.

My own "The Conclusion of Practical Reason" revisits the question of what is properly considered the conclusion of practical reasoning. Aristotle famously argued that the conclusion of practical reasoning is an action, but contemporary work in practical reason seems to find this Aristotelian thesis puzzling. After all, it seems that I may go through a perfectly complete piece of reasoning, form an intention, and never act on it. It seems that intentions have a better claim to be the conclusion of practical reasoning than actions. However, I argue that suitably understood the Aristotelian thesis is correct: only actions can be a proper "resting place" for practical reasoning.

Carla Bagnoli's "Phenomenology of the Aftermath: Ethical Theory and the Intelligibility of Moral Experience" examines the relation between ethical theories and moral phenomenology, in particular the question of whether ethical theories are necessarily in tension with our moral experience in cases of conflict. The unavoidable experiencing of emotions such as regret in cases of conflict seems to be left unaccounted by any ethical theory that will resolve the conflict by selecting one option as the ultimately correct one. Given this possible tension, it seems that one has two options: either one takes the moral theory to provide the standard by which we judge the appropriateness of our experience, or we take our experience to provide criteria of adequacy for moral theories (leaving open the possibility that no ethical theory will be prove to be

adequate). However Bagnoli argues that this limited range of options is predicated on an incorrect understanding of the role of ethical theorizing, a view that takes the demand for ethical theorizing to be external to our practical concerns. Bagnoli argues that, by presenting us with a moral ideal, ethical theory contributes in a distinctive way to our understanding and articulation of our moral experience. On the other hand, ethical theory is constrained by moral experience by having to make the agent's experience intelligible to her.

The topic of the significance of our experience of regret is further explored in Connie Rosati's "Mortality, Agency, and Regret." Rosati examines what the experience of regret tells us about the nature of our good. Rosati argues that the possibility of well-grounded regret is a consequence of what she calls "the circumstances of the good," the objective and subjective constraints that our agency faces that limit in important ways the pursuit of the various things we desire. Rosati argues that given these circumstances of the good, the same capacities that render us autonomous will lead us to experience regret. Rosati argues that this analysis of regret throws important light in our understanding of the nature of the good, in particular, in understanding the need of forming, and the constraints governing, one's conception of one's good.

Lastly a small word of clarification. This volume has been many years in the making, and Phil Clark and Donald Ainslie were invited very early in the process. At the time I invited Donald to write for the volume I was teaching at the University of New Mexico, and at the time I invited Phil he was teaching at Kansas State University. We all ended up at Toronto, and the only unfortunate consequence of having such terrific colleagues is that I ended up editing a volume with three contributors from the same institution.

University of Toronto
Department of Philosophy
215 Huron St.
Toronto, M5S 1A1, Canada
e-mail: sergio.tenenbaum@utoronto.ca

James Doyle

DESIRE, POWER AND THE GOOD IN PLATO'S *GORGIAS*

An ancient philosophical saying has it that everyone desires the good or, more precisely, that everything pursued is pursued as being something good (*omne appetitum appetitur sub specie boni*). At first sight, this may seem to be empty or false depending on how it is interpreted. If by 'the good' we mean something entirely formal, such as whatever is desired or is the highest object of desire, the saying merely records that everyone desires what they desire. But if we mean something more substantive by 'the good', such as the Good (Plato), virtuous activity of the soul (Aristotle), intellectual contemplation (Aristotle) or pleasure (Epicurus, Mill), the saying is false, since there seem to be people who desire none of these things.

A more promising interpretation is that everyone, insofar as he is rational, desires his real (as opposed to apparent) good, whatever that may be. This is not empty, as it represents a substantive constraint on what is to count as rational desire and action. And it seems likely to be true, since if there is such a thing as practical irrationality, acting against one's acknowledged real good is surely a case of it. Nevertheless there are many philosophers who would deny that it is true, or that it tells the whole truth, even on this interpretation. Moral psychologists in the tradition of Hobbes and Hume, for example, would be likely to object to the implication that there is such a thing as one's real, as opposed to apparent, good, on which any philosophical weight may be placed. If we are to speak of our good at all, they will object, we should understand it as supervening upon our desires; and if there is any room for a conception of our real good as opposed to our apparent good, it should be understood entirely in terms of what we would desire upon reflection if

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 15-36. Amsterdam/New York, NY: Rodopi, 2007.

we had all the relevant information, as opposed to what we desire on the spur of the moment or from a position of debilitating ignorance.[1]

Nihilism about the real good is, in a way, all very well: it is perfectly intelligible to claim that, contrary to appearances, there is no such thing as an agent's real good, and no such thing, really, as a reason to act. (I believe that this is what Hume's view amounts to.) But anyone who suggests that a philosophically viable facsimile of the real good can be constructed out of the materials available to the anti-realist about the good can hardly be encouraged by the history of modern epistemology. Phenomenalism was an attempt to reach an accommodation with scepticism about the external world, by constructing a facsimile of that world out of brute appearances or sense-data. But one hears very little about sense-data nowadays, and very many philosophers doubt that the idea of appearances makes proper sense once it is cut loose from anything they could be appearances *of*.[2] Desires, "appearances" of a nonexistent good, are then the sense-data of the practical realm, and the challenge facing the "soft" or reconstructive antirealist about the good is structurally the same as the one the phenomenalists spectacularly failed to meet. The facsimile of the real good must, if the theory is to save the basic appearances, explain the illusion to which we are subject when we are conscious of acting as if in pursuit of our real good. And if practical nihilism is to be avoided, there is one aspect of that experience that cannot be dismissed as an illusion at all: the *rationalizing force* that its directedness toward our real good bestows upon our action. But if our desires can't count as providing reasons for action in virtue of disclosing an aspect of our real good to us, how else could they provide such reasons? What's so good, in itself, about getting what you want?

The very idea of sense-data was, arguably, one of the many made possible by Descartes; it is unknown to the ancients (see Burnyeat 1982). I suspect that its equivalent in the practical realm would have been likewise untintelligible to them. It certainly would have been unintelligible to Plato, who deploys in the *Gorgias*, I will argue, the constraint on rationality mentioned above, that rational agents desire and pursue their real good. As with sense-data, which bear no necessary relation to any entity beyond themselves, there is a tendency these days to think of the "apparent good," where this is given to us in desire, as ontologically

[1] A conception of desires as "appearances of the good" is articulated in Tenenbaum (2007); see also Tenenbaum (1999).

[2] The earliest example of this line of argument is probably Arnauld's critique of Malebranche in his 1683; Reid (2002) was also influential. More recently, see Austin (1964) and McDowell (1994).

independent of the "real good," so that it can legitimately be enquired whether the latter exists even once the reality (such as it is) of the former has been conceded. This would have made no sense to Plato because his conception of the apparent good just is *the way the real good appears*, the idea of appearance presupposing the existence of a reality.

This way of conceiving the relation between real and apparent good is presupposed in the argument that Plato has Socrates put to Polus in the *Gorgias*, for the claim that tyrants have no real power and, more generally, that no-one who acts unjustly does what he wants. The argument has two stages. First, Socrates tries to establish that no-one who acts against their own interest does what they really want. Then he tries to show that anyone who acts unjustly acts against their own interest. My concern here is with the first stage. My claim is that, given some reconstructive surgery, it is valid. Its premises, which include Plato's conceptions of real and apparent good, are contestible; but those who regard as sisyphean the "reconstructive" attempts to do justice to what we know about what being an agent is like, and look upon nihilism about the good as a counsel of despair, will have reason to hope that the argument is also sound.

Before scrutinising the argument Socrates puts to Polus about desire and the good, I would like to say something about how it serves the philosophical purposes of the *Gorgias* as a whole. We are usually told that the dialogue's main themes are rhetoric and ethics (where ethics and politics are not distinct, as they were not for the Greeks generally). If we really think about this platitude, however, we are likely to be puzzled, because we are not nowadays used to the idea that rhetoric and ethics have much to do with each other. But if we think of the *Gorgias*'s themes as *persuasion* on the one hand and *politics in its ethical dimension* on the other, we are at the same time more faithful to the dialogue and better placed to see how its themes are inseparable. The common conception of ethics as a kind of law tends to make us forget that a large proportion of the misery human beings deliberately inflict upon each other does not violate but is permitted, and may even be required, by the positive law or edicts of the relevant local regime. A great attraction of political power has always been the opportunities it affords to disregard others' interests with impunity. The most serious threat to ethical life does not come from Hume's knave or Hobbes's Foole, furtively pursuing an egoistic agenda while putting on a show of moral probity. It comes from those who seek the natural ring of Gyges: the kind of political power that will free them entirely from the constraints of ordinary morality. This permanent ethical danger is part of the essence of politics. But the acquisition of political

power is always partly a matter of persuasion, even where its possession can be sustained only by force. So anyone going in for wrongdoing on the grandest scale will be interested in the dynamics of persuasion. Rhetoric features in the *Gorgias* only because it constituted the art of persuasion in Plato's world. None of the dialogue's arguments turn on the fact that persuasion in that world was effected by rhetorical tropes, as opposed to 'branding', sound-bites, spin, advertising, or any other components of the apparatus of political manipulation that have replaced rhetoric in our own time (see Doyle 2006a).

The overarching aim of the *Gorgias* is to denounce as unworthy the acquisition of political power as an end in itself, and to expose as corrupt the art of persuasion, when it is cut free from any concern with truth, knowledge or the good of the persuaded. Since Socrates believes that a concern for virtue, of the kind that the usurpation of political power apparently enables one to dispense with, is always in the agent's own interest, we should expect him to argue that tyrants have no *real* power – that is, power understood as necessarily beneficial to its possessor.

1. Can Socrates Be Serious?

The claim that tyrants have no real power certainly has a paradoxical ring to it – particularly if one refuses to distinguish between political power and real power. Some commentators believe that Socrates cannot be serious in advancing it (e.g. McTighe 1992, esp. pp. 280-281; Weiss 1985 and 1992). Among their reasons, they emphasize the intrinsic implausibility of Socrates' arguments. This aspect of their case must await our scrutiny of his claims' philosophical credentials below. But there is also the matter of extrinsic considerations. These suggest that Socrates means exactly what he says.

1. The sceptics maintain that there is something about Polus in particular which explains why Socrates would be satisfied with a merely *ad hominem* refutation of him. Citing the "purgative dialectic" outlined at *Sophist* 230a5-d4, Kevin McTighe writes:

> He who is firmly convinced of his wisdom is induced to contradict himself solely in order for him to confront and acknowledge his actual ignorance [. . .].
>
> [. . .] [S]uch a one as Polus, conceited to an extreme, is unable and unfit to take part in [. . .] [a sincere] [. . .] investigation. The character of the dialectic changes accordingly. Since Socrates is now arguing not for the

> sake of the *logos* but for the sake of Polus, he has no reason necessarily to adhere to the standards of logic or fair play – and possibly reason not to. To achieve his ends, the practitioner of the purgative method may find it simply more efficient to argue fallaciously, to admit into discussion theses he regards as false [. . .].
>
> [. . .] Socrates succeeds in getting [Polus] to deny certain assumptions [. . .] which at first seem self-evident: [. . .]. It doesn't matter that these assumptions are perfectly legitimate. Nor that the denial of [the first] rests on grounds which Socrates himself does not accept [. . .]. (McTighe 1992, p. 280)

But how much significance can we accord to the fact that Socrates happens to be addressing Polus? It is true that Socrates' manner changes markedly with each interlocutor: conventional courtesy masking his occasional mockery of Gorgias; sarcasm and superb disdain toward Polus; and genuine respect for Callicles giving way to genuine dismay. But this should not cause us to forget that the whole sequence of conversations is conducted in front of a sizable audience.[3] Socrates cannot therefore be tailoring his conversation exclusively to any one of his interlocutors, or even to all three of them. Before such an audience, his changes of tone must be understood at least in part as for the audience's consumption. So if he is saying things to Polus that he doesn't believe, this cannot be entirely because he thinks that he is thereby benefiting Polus: he must also conceive of his deceit as having a point for his audience: either he is taking them in as well, or he is winking at them behind Polus' back. The audience, which originally came to hear Gorgias's rhetorical display, is presumably made up largely of Athenian intellectuals. If the sceptics are right that Plato expects his acute readers to realise that Socrates is deceiving Polus, Socrates must expect his acute hearers to realise the same thing. The question the sceptics must then confront is: *what would these members of the audience then be expected to make of Socrates' prior and subsequent attacks on rhetoric for its lack of concern with knowledge and the truth* (459b6-c2, 462c3-7,

[3] Since Socrates and Chaerephon arrive immediately after the end of Gorgias's display of rhetorical virtuosity. Plato does not emphasize this. The continued presence of the audience doesn't become apparent until 458c3, where Chaerephon responds to the suggestion that the discussion should perhaps end there by saying, "You can hear for yourselves from their clamoring, Gorgias and Socrates, that these men want to hear anything you say." When Socrates asks Polus whether his laughter at the doctrine that it is always worse to commit injustice than to suffer it is "a new kind of refutation: [. . .] to laugh, instead of refuting," Polus replies, "Don't you think you've been thoroughly refuted, Socrates, when you say the kind of thing that no-one else would say? But then, ask any of the people here" (473e4-5).

500e3-501c6, etc.)?[4] If Socrates is deceiving Polus, he intends to be understood as doing so, in which case he is deliberately making a mockery of his own critique of rhetoric.

2. The sceptics argue that it is "silly" or otherwise implausible for Socrates to claim that power benefits its possessor (see McTighe 1992, p. 277; Weiss 1992). But it is not even *prima facie* odd for Socrates to make this claim. The sceptics presuppose that Socrates is talking about what people conventionally understand by 'power'. And there is more generally a curious reluctance among commentators to concede that he may use key terms in a "special sense." Yet he does this constantly in the *Gorgias*. If he is working with a conception of *real* power, which is by definition good for its possessor, this fits perfectly well with his appeals throughout the dialogue to *real* notions as distinguished from their conventional counterparts: *real* wanting, which necessarily takes as its object the agent's *real* good; people's *real* preferences, which are in fact for suffering injustice over committing it (474, 475e); Polus's *real* beliefs, among which is to be numbered that tyrants and orators have no power (466e); *real* happiness, which depends entirely on "education and virtue" (470e); *real* proof, which is concerned to persuade only one's interlocutor, whatever anyone else may believe (472); *real* nobility and goodness, which take no thought for the length of one's life (512d-e); the *real* art of statesmanship, of which Socrates, of all people, may well be the only student (521d); and finally *real* oratory, which has as its object the edification of its audience (503a, 504d, 508c). In none of these cases is the distinguishing mark of the real notion any less offensive to conventional wisdom than the claim that real power benefits its possessor. Consistency would drive the sceptics to the absurd conclusion that Socrates is dissimulating throughout.

3. Socrates tells Polus that the issue of who is happy and who not is "that about which knowledge is most honourable and ignorance most shameful" (472c); the issue of whether tyrants have power, where power is a *good*, is part of this subject. Socrates also accuses Polus of trying to eject him from "[his] lawful property, *the truth*" (472b). He denounces Polus's method of refutation, which consists in pointing out that Socrates is saying things that no reputable person would agree with, as useless in establishing *the truth* (471e). Of course it is often notoriously difficult to establish when Socrates is being serious. But as in the *Apology* (Plato

[4] References to the *Gorgias* are by Stephanus numbers from the text in Dodds (1959); to other Platonic dialogues by Stephanus numbers from Plato (1963). All translations are my own.

17b5-6, c1, 18a5-6, 20d4-6), so in the conversation with Polus: a reader who denies that Socrates is serious must explain why he insists with such emphasis that his overriding concern is with the truth.

Aside from all this, and from the question of how *plausible* in itself it may be to assert that tyrants have no real power and do not do what they (really) want, there is reason to think of these as sayings we can intelligibly *hope* to be true. Polus and, especially, Callicles are obviously very much taken with the glamorous allure of political power. Callicles' position, as set out in his great speech, is a subtle and compelling one, brilliantly expressed. Unlike the modern "moral sceptic," a merely notional figure, Callicles articulates a way of life which is not only a real option, but was a lived reality among many members of the political class to which he belonged, and had its roots in the heroic tradition with which all educated Greeks were imbued. There is nothing anomalous in his having inspired Nietzsche – not a superficial man – nor anything fantastic in Dodds' conjecture that he represents Plato's conception of what he might himself have become had he not quit politics.[5] Polus is in thrall to as much of Callicles' ideology as he understands (not a great deal), and for Socrates to tell Polus that tyrants actually have *no power* is to hit him where it hurts. If power is worth seeking, it makes perfect sense to hope that it is not, after all, what tyrants have found, so that one can neutralise the attraction political power holds for people like Polus by telling no more than the truth about tyrannical impotence. This insight into the glamour of political power pervades the *Gorgias*, and it shows that Plato is unusual among moralists in refusing to assimilate all vices to what is mean, conniving and parasitic in human beings. Nor is it easy to see how the *Gorgias*'s conception of moral danger can be made consistent with the assimilation of all vice to ignorance which Socrates seems to insist upon in many of the shorter dialogues.

2. The Argument: Translation and Commentary

Socrates has questioned Gorgias, the pre-eminent teacher of rhetoric, about what his *technē* (art, craft, science, skill) consists of; and he has made Gorgias give the appearance, at least, of contradicting himself. Polus, Gorgias's impetuous acolyte, then takes the reins and tries to get Socrates to say what kind of *technē* he thinks rhetoric is. After expressing

[5] Dodds (1959, Introduction, p. 14); see also Dodds (1959, Appendix, pp. 387-391) on Callicles and Nietzsche.

apprehension that his answer may offend Gorgias, Socrates replies that he doesn't think that it is a *technē* at all, but merely the knack, acquired by experience, of flattering and giving pleasure to the soul without regard to the soul's real good, much as cookery aims at giving pleasure to the body without regard to health. Polus, nonplussed, asks Socrates whether "the good orators are considered base flatterers in the cities"; to which Socrates' odd reply is that he thinks that they are "not considered [*or* esteemed] at all" (οὐδὲ νομίζεσθαι ἔμοιγε δοκοῦσιν). Polus is now utterly perplexed:

> *Polus:* Don't they have great power in their cities?
> *Socrates:* No; not if you mean by power something that benefits the powerful.
> *Polus:* But I do mean that.
> *Socrates:* In that case, I think that orators have the least power of anyone in the city.
> *Polus:* What? Don't they kill whomever they want, just like tyrants, and banish people and confiscate their property whenever it seems good to them to do so? (466b4-c2)

Orators can be assimilated to tyrants for the purposes of this discussion because tyranny is the natural *telos* of rhetoric. Since the skilful orator can persuade anyone to do whatever he wants, it is only a matter of time before he takes control of the city; Gorgias's own eulogy of rhetoric explicitly claimed that it enables its practitioner to enslave the practitioners of other *technai* (δοῦλον μὲν ἕξεις τὸν ἰατρόν κτλ, 452e1-8). Why does Polus latch onto the ability to kill, banish and confiscate? Presumably because

(i) he thereby inadvertently reveals himself as a power-worshipper of a particular kind: the aspect of power to be emphasized in a presentation of it as desirable is its licence to dominate others and not, for example, its providing the opportunity to build more libraries or alleviate poverty. (This is of a piece with his conception of power as something glamorous, and will be important later on. Callicles' power-worship will prove to be of the same kind, but Callicles' unveiling of it will be anything but inadvertent.)

(ii) These components of power are not only its particularly apt representatives when it is understood in this way as the ability to bend others to your will; they can also be seen as its basis: the credible threat of killing, confiscating or banishing goes a long

way toward enabling the tyrant to get others to do pretty much whatever he wants.[6]

Socrates claims that Polus's last question is actually two questions, because doing what one wants or wills is not the same thing as doing what seems best to one:

Socrates: I maintain, Polus, that orators and tyrants alike have the least power in their cities [. . .] because they do nothing that they want (so to speak); although they certainly do what seems best to them.
Polus: Then surely this is great power?
Socrates: Not according to Polus, no.
Polus: Not according to me? But according to me it certainly is.
Socrates: By the . . . *Not* according to you, seeing that you say that great power is a good thing for the one who has it.
Polus: Well, yes; I do say that.
Socrates: So you think it's a good thing, do you, for someone to do what seems good to him, even though he has no understanding [νοῦν μὴ ἔχων]?
Polus: I do not.
Socrates: Then presumably you'll show me that orators do have understanding, by refuting my claim that rhetoric is not a *technē* but merely flattery? If you let me go unrefuted, orators and tyrants who do what seems good to them in their cities don't thereby have access to anything good. But power is, as you say, a good thing; whereas even you would agree that doing what seems good to you without understanding is a bad thing. Wouldn't you?
Polus: I would. (466d6-467a6)

It's important to keep track of the state of dialectical play here. First, the apparently arcane and typically Socratic subject of the earlier conversation with Gorgias – whether rhetoric is a *technē* – turns out to be crucially relevant to the new subject taken up by Polus – whether orators have power. If power is a good thing; and the orators' claim to power consists in their doing what seems good to them; but doing what seems good to one without understanding is not a good thing; and the orators practice no *technē* and so act without understanding; *then* it follows, according to Socrates, that the orators have no power. (Of course, it only

[6] A third reason may be that these basic and representative components of power allude to Socrates' later trial and condemnation, in that these were the penalties among which Socrates was expected to choose. Certainly the later conversation with Callicles reverberates with unmistakeable pre-echoes of Socrates' death, some of them very bitter.

follows if the orators' lack of a *technē qua* orators disqualifies them from acting with understanding in doing what seems good to them, and it's hard to see why this should generally be true. Why should the fact that the orator (and the cook) do not earn their living by practicing a *technē*, but merely by exercising a knack, entail that they lack understanding in all their activities? How can their not knowing the *technai* of cookery or rhetoric tell against their general understanding, *particularly* if there are no such *technai* to be known? The silence of the *Gorgias* on these questions seems to me a serious *lacuna* in its main line of argument.)[7]

Secondly, since Polus immediately agrees, the issue about whether orators have power is effectively over already. The burden of proof is squarely upon Polus, to show that the orators act with understanding. But the issue of doing what one wills or wants (ἃ βούλεταί τις) *versus* doing what seems good to one (ἃ δοκεῖ τινι) has barely been broached. It follows that this latter distinction is irrelevant to the debate about the orators' power. This is interesting, because Socrates later talks as though he has established the orators' impotence[8] by means of his distinction between doing what one wants and doing what seems good:

> So I was speaking the truth, when I said that it's possible for a man who does what seems good to him in the city not to have great power or to do what he wills. (468e3-5)

In fact, he has already established it here, two and a half Stephanus pages earlier. So Plato cannot have had Socrates introduce the distinction in order to support the claim that tyrants have no power. Therefore even if the sceptics (see McTighe 1992; Weiss 1985 and 1992) were right to suppose that Socrates did not believe this claim and argued for it only to humiliate Polus, it would not follow that he is not serious about the distinction, or about the claim which, as we shall shortly see, the distinction enables him to formulate: that those who do, without intelligence, what seems good to them do not thereby do what they want.

Polus, now at a loss, agrees to let Socrates ask the questions, to make his meaning clearer.

---

[7] Penner, in (1991, pp. 157-165), tries to fill the *lacuna* by arguing that Gorgias and Polus see mastery of rhetoric as rendering the study of any other *technē* unnecessary: the orator is better than any specialist at persuading the many about matters that belong to his specialisation; he can therefore "enslave" the practitioners of all other *technai* (452e4-8). But this does not show why the orator *as such* must lack knowledge of any other *technē*; Gorgias and Polus are expressing an attitude towards rhetoric, not evincing an attitude it requires.

[8] Strictly speaking: the compatibility of the orators' impotence with their doing what seems best to them.

Socrates: Do you think that men want[9] whatever they do on any given occasion, or that for the sake of which they do what they do? For example, those who drink medicines on their doctor's advice: do you think that they want what they are actually doing, drinking the medicine and suffering pain, or rather being healthy, for the sake of which they are drinking it?

Polus: Obviously they want to be healthy.

Socrates: I presume that people who go on sea-voyages or engage in other kinds of trade don't want this, what they're doing at the time (I mean, who wants to go on a voyage and put himself in danger and increase his troubles?), but I think they want that for the sake of which they go on the voyage: making money; because they take the voyage for the sake of money.

Polus: Of course.

Socrates: Isn't this how it is with everything? If someone does something for the sake of something else, he doesn't want what he does, but that for the sake of which he does it?

Polus: Yes.

Socrates: Well then, is there anything in the world that isn't either good, bad, or between these two – neither good nor bad?

Polus: Necessarily, everything must fall under one of these heads, Socrates.

Socrates: No doubt you'd say that wisdom, health, wealth and other such things are good, and their opposites bad.

(At this stage Socrates is considering action for the sake of an end from the point of view of the agent; he is trying to establish that we want ends rather than means when action is viewed "from the inside." So nothing really turns on whether Socrates himself would endorse this list of intrinsic goods.)[10]

Polus: I would.

Socrates: And would you say that things like this are neither good nor bad: things that sometimes share in the good, sometimes in the

[9] βούλεσθαι, 467c5. I avoid 'will' because (i) it is stylistically awkward at most points in this passage, and (ii) since will tends to be opposed to desire, particularly in post-Kantian philosophy, while βούλεσθαι in Greek is much more versatile, this translation would give unearned plausibility to my view that Socrates intends his βούλεσθαι/δοκεῖν τινι distinction seriously, and would make Polus's immediate denunciation of it as "shocking and monstrous" (Σχέτλιά γε λέγεις καὶ ὑπερφυῆ, 467b10) close to inexplicable. McTighe (1992) translates 'desire', which strikes me as tendentious in the opposite direction; it has the advantage, for his thesis that Socrates does *not* intend the distinction seriously, of making Socrates talk nonsense from the moment it is introduced.

[10] In any case, I'm not aware of anything in the *Gorgias* that rules out Socrates' believing that health and wealth are intrinsic goods *when they are accompanied by virtue or wisdom*; this seems to be the view in Plato's *Apology* (30 b2-4); cf. Plato (278e2-281b6).

bad, and sometimes in neither; sitting, walking, running and sailing, and then things like stone and wood and the like? Wouldn't you say that? Or would you describe anything else as neither good nor bad?

Polus: No, that's right.

Socrates: Now, when people do these in-between actions, do they do them for the sake of good things, or is it the other way around?

Polus: They do the in-between actions for the sake of good things, of course.

Socrates: So it's in pursuit of what's good that we walk, when we walk, thinking that it's better to do so; and on the other hand, when we stand still, we stand still for the sake of the same thing: what's good. Or isn't that right?

Polus: Yes, that's right.

Socrates: And we also kill – if we're killing someone – and banish, and confiscate property, in the belief that it's better for us to do these things than not to do them?

Polus: Of course.

Socrates: Then men do all these things for the sake of the good?

Polus: I should say so.

Socrates: Didn't we agree that when we do something for the sake of something else, we don't want this first thing, but that for the sake of which we do it?

Polus: Absolutely.

Socrates: Then we don't want to slaughter or banish from cities or confiscate property simply as such, but we want to do them if they are advantageous,[11] and don't want them if they are harmful. For we want good things, as you say, and we don't want what's neither good nor bad, nor what's bad. Well? Do you think I'm speaking the truth, Polus, or not? Why don't you answer?

Polus: You're speaking the truth.

Socrates: So if we're agreed upon this, then if someone, whether he's a tyrant or an orator, kills someone or banishes him from the city or confiscates his property, in the belief that it's better for him to do so, although it happens to be worse, this man is doing what seems good to him. Right?

Polus: Yes.

---

[11] Superficially Socrates here contradicts what he said earlier about a man's *not* wanting, e.g., the dangerous sea-voyage but the good that comes of it; clearly he was there talking about wanting something *for its own sake*; nb. that this is not the same as his conception of "*real* wanting": an action (e.g. banishing) can be an object of a "real want" if it leads to good without being itself a good. See Penner (1991, p. 179).

Socrates: Then does he do what he wants, if these things happen to be bad? Why don't you answer?
Polus: I don't think he does what he wants.
Socrates: Then is there any way that this man can have great power in this city, if, as you agreed, great power is something good?
Polus: There's no way.
Socrates: Then I was speaking the truth, when I said that it's possible for a man who does what seems good to him in the city not to have great power or to do what he wants. (467c5-468e5)

What does this distinction, on which Socrates' argument turns, between what one wills or wants (ὃ βούλεταί τις) and what seems good to one (ὅ τινί δοκεῖ [ἀγαθόν]) amount to? It's hard to escape the conclusion that it amounts to a distinction between different senses of 'want'. He can hardly be claiming that there's no sense in which one wants to do what seems good to one when one does it because it seems good to one. Commentators have pointed out (e.g. Weiss 1992, pp. 308-309) that he raises no objection later (511a5-7) when Callicles uses βούλεσθαι in just this sense. There's a well-established sense of *want* (and of βούλεσθαι) according to which wanting to do what one does is a necessary condition of doing it freely or intentionally at all. Socrates cannot mean that the tyrant's actions are to be classed with unintentional or inadvertent acts, because he thinks that the tyrant is responsible for them, in that he deserves to be punished for them (476a2-481b5, 525b1 *ad fin.*). Socrates seems to mean something like this: the tyrant's actions don't in fact conduce to what he *really* or *ultimately* wants; *viz.* his own good. "What seems good to one" is one's apparent good, or the (possibly distorted) way one's real good appears, and there is a well-understood sense in which this is what one wants; the only thing left for "what one (really) wants or wills" to mean is one's real good, whatever that should turn out to be.[12]

It has been suggested (Segvic 2001) that 'Socratic wanting' in this passage requires that the wanter *know* that the object of their want is (or

[12] Penner (1991) denies that Socrates deploys a "special sense" of βούλεσθαι, and seems to think that this requires him to attribute to him instead the view that the identity of an action depends upon its consequences, however distant (§11, pp. 185-192). But such a view seems to entail that none of us ever knows what he is doing. Nor is there any evidence in the text for the attribution. Penner himself concedes that Socrates' argument goes through without it, but in a way that "leaves it as an undefended premiss that power is always good for its possessor" (p. 190). But once we understand Socrates' argument, we can see why Polus has to accept that power is always good. For Polus conspicuously *wants* power; and it is one of the main lessons of the argument that we can only reflectively endorse our wants if we can conscientiously affirm that their objects are good.

will lead to) good. But nothing in the passage warrants this interpretation. By this I mean:

(i) Socrates never says that this is a condition on "real wanting"; the only condition mentioned is that the object *be* good;
(ii) if the argument for tyrannical impotence goes through, it goes through without any need to appeal to the wanter's knowledge, and if it doesn't, appeal to the wanter's knowledge is not going to help.

The only contribution such an appeal could make to the argument is via the entailment, from something's being known, that it is true; but the relevant truth – that the object is good – is already provided for in the argument as Plato presents it. Certainly we may presume that only a person with justified – in the currently popular sense of *reliable* – beliefs about which actions and consequences are likely to be good is going to do much by way of real wanting. But we may justly suspect that Socrates cannot envisage that anyone who really wants something must *know* it to be good; otherwise real wanting may turn out to be humanly impossible. It is an important message of Plato's *Apology* that knowledge of anything "fine and good" (καλὰ κἀγαθά) is not to be had by human beings. Knowledge of what sorts of things are good must certainly itself qualify as "fine and good," and the impossibility of such knowledge is strongly implied in the *Gorgias* too, in the famous "iron and adamant" passage (508e6-509a7), in which Socrates disavows knowledge of the fundamental ethical principles of which he is seeking to persuade Callicles: "With me *it's always the same story*: I don't know how these things are [. . .]" (a4-5).

What more can we say about the relationship between the desires for the real and apparent good? As I began by saying, everyone desires the (their) true good in the following sense: if they discover that the good that they've been pursuing is *merely* apparent (i.e. not a good at all), they no longer desire to pursue it (unless they're weakwilled or otherwise irrational). People don't desire the apparent good *as such*; this is why desires don't, as such, provide reasons for action.

But of course Socrates is saying more than this. On his view, if one's apparent good is not in fact one's real good, *one doesn't really desire it*. I have argued above from considerations extrinsic to 466-468 that we lack good reason to deny that Socrates is speaking *in propria persona* when he maintains against Polus that tyrants and orators have no power and don't do what they want. The key to defusing the air of paradox and absurdity surrounding both claims – and thereby removing the main obstacle to taking Socrates at his word – lies in the realisation that he is

talking about *real* – that is, beneficial – power and what one *really* – that is, unconditionally – wants, where these are explicitly contrasted with their standardly-understood counterparts.

This gets missed, it seems to me, if we construe Socrates' concept of "wanting" (βούλεσθαι) in this passage after the fashion of a "technical" notion or term of art, of the sort that philosophers introduce from time to time as a means of abbreviating their argument or avoiding certain complications.[13] This ignores the asymmetry between Socrates' concept and what the many understand by the same word, and leaves us wondering how the claim that wanting always takes the real good as its object can have the philosophical significance Socrates clearly intends for it, if "wanting" here is merely an idiosyncratic philosophical construct.[14] Socrates clearly thinks that his notion corresponds to what wanting *really* is, or (equivalently) what the *real* object of wanting is; the many *conflate* this idea with that of one's being subject to an *appearance of good*, which is somehow a "counterfeit" of wanting. What justifies Socrates' viewing his conception as the real one? As we shall see below, Socrates' conception is of wanting *as constrained by reason*, so that what one wants can be reflectively endorsed as liable to contribute to (one's) good, or at least not detract from it. One reason for thinking of such a conception as what wanting *really* is would be a conviction that *a rational being* is what a human being really is: Socrates' conception of what one *really* wants is simply what a rational being *would* want. He seems to express just such a conviction in Plato's *Apology*. For there he says that incomparably the most important thing a human being can do, and which every human being has a natural[15] duty to do, is to subject his way of life to rational scrutiny. *The unexamined life is not worth living for a human being*. This account of the human essence pretty obviously runs through the *Gorgias* as well.[16]

---

[13] This is how I understand Segvic (2001); she refers to Socrates' concept as "Socratic wanting" (p. 9), and describes Socrates as "maintaining that we want – *in some legitimate sense of this word* – what as a matter of fact is good" (p. 45, my emphasis).

[14] "[W]hy should [Socrates] construe 'wanting' in such a peculiar way?" (Segvic 2001, p. 6). If we understand Socrates as making a claim about *what wanting really is*, this question doesn't arise. (This is not to deny that a similar-looking question may still be raised by someone who finds Socrates' conception of *what wanting really is* peculiar.)

[15] As I understand Socrates' position in the *Apology*, while his own special duty to exhort his fellows to self-examination had a supernatural origin in "the god" speaking through the Delphic oracle, the duty to self-examination itself, to which all human beings are subject, is no divine decree but is imposed by the conditions of human life. See Doyle (2004).

[16] See especially 472c6-d1 and 500b5-c8. The *Gorgias* seems to specially invoke the *Apology* in a number of places. Its very first lines allude to the *Apology:* told by Callicles

3. Reconstructing the Argument

While Socrates' argument is not so implausible that we must doubt his seriousness, it will not do as it stands.

First, there is the distinct suggestion that every action is done as a means to the (agent's) good. It is true that Socrates does not say this in so many words; but he does say it of all bad or indifferent actions (468a5-c7), and he gives no examples of actions good in themselves, as opposed to states like health and wisdom (467e4-5). Certainly commentators have not been slow to attribute the suggestion to him, and even to endorse it.[17] But it is a false and pernicious bit of philosophy. Sometimes we simply do what we do. Consider doodling while talking on the telephone. This may well be an action: deliberate, conscious, intentional, not inadvertent, etc. Yet, even when a fullblown action, it need not be done with any end in view beyond itself. It is pretty clear that any attempt to make it out as done for the sake, say, of diverting or entertaining oneself can only be born of a prior conviction that there *must* be some distinct end: in the absence of any such conviction, no-one would dream of talking this way (see Anscombe 1963, §§ 17, 21).

Since it would be absurd to denounce doodling, when done in this spirit, as irrational, this entails a second, distinct point: to say that (we must reasonably believe that) an action contributes to our good if it, or the desire to perform it, is to count as rational is to express too strict a conception of rationality. At most, an action and the desire to perform it fails of rationality if (we reasonably believe that) it detracts from our good.

Finally and most importantly, Socrates' preferred distinction of what we really want from what merely seems good to us makes all of our real desires except for our desire for our real good conditional, in a way that makes it unknowable what will satisfy them. (For the sake of simplicity I for now ignore the previous objections.) As we have seen, Socrates takes his examples of actions bad or neutral in themselves (taking medicine, walking, killing, banishing etc) to establish his general conclusion: "Then

---

that he has just missed a dazzling rhetorical display by Gorgias, Socrates replies "This is all Chaerephon's fault, Callicles; he made us linger in the market-place." It was Chaerephon's question of the Delphic oracle, of course, that led Socrates to spend most of his subsequent life in the market-place, as he tells the jury in the *Apology*. See Doyle (2006b).

[17] "[. . .] I believe Socrates would argue that *no* action is ever undertaken for its own sake. I even believe Socrates would be right to argue that" – Penner (1991, p. 185) (emphasis Penner's). Given Socrates' unequivocal statement that where there is a distinct end, it is always the good, this amounts to attributing the suggestion to him.

we don't want to slaughter or banish from cities or confiscate property simply as such, but we want to do them if they are advantageous, and don't want them if they are harmful. For we want good things, as you say, and we don't want what's neither good nor bad, nor what's bad" (468c2-7). Notice that he says "if they *are* advantageous" (ἐὰν μὲν ὠφέλιμα ᾖ) and "if they *are* harmful" (βλαβερὰ δὲ ὄντα) and not "if they *seem to be* [. . .]" or "if we *believe them to be* [. . .]," because he is contrasting what we really want with what (merely) seems good to us. Now, suppose I desire to φ,[18] where φ-ing is some action within my power. Let *A* be the proposition that I φ, and let *Å*[19] be the proposition that my φ-ing, and so the truth of *A*, contributes to my real good. According to Socrates:

R1 If *Å*, then my desire is satisfied by my φ-ing.
R2 If ~*Å*, then my desire is satisfied by my not φ-ing.

But I can never know which of *Å* and ~*Å* is true. It is part of the human predicament that if I really know that something will contribute to my real good, I don't know how to get it, because the only such thing is *my real good whatever that should turn out to be*; whereas if I know how to get something, I don't know that it will in fact contribute to my real good. Even if I've thought the matter through as carefully as possible, I cannot know that bad luck won't result in my subjectively rational action having a disastrous consequence, as when a fit jogger suffers a freak heart attack.[20] It follows that, unless my desire is for my real good, in which case *Å* is necessary and knowable *a priori*, *I don't know what will satisfy any of my real desires*. The notion of a real or rational desire then becomes entirely useless in any account of what I do. Even if this consequence is not strictly intolerable, we should avoid if at all possible an account which makes us so unintelligible to ourselves. Yet our

[18] That is, I am in a state which would normally be characterized as desiring to φ.

[19] The ingenious notation is David Lewis's; see his (1988, p. 326); apparently *Å* is pronounced "A-halo"; see Price (1989, p. 122).

[20] But notice that these possibilities are not freakish in a way analogous to, say, the brain-in-a-vat hypothesis. Aristotle, who identifies an agent's good with his happiness, notes that human life is ineliminably uncertain, in such a way that any plausible account of happiness must include an element of good fortune which it is beyond the agent's power to guarantee – as it was beyond Priam's (Aristotle 1963, 1100b23-1101a14). Arguments that the BIV possibility is too freakish to defeat my claim to know things which presuppose that I am not a BIV (such as that I have two hands) therefore cannot be converted into arguments that the possibility that my subjectively rational action may result in disaster is too freakish to defeat my claim to know things which presuppose that it won't (such as that it will contribute to my real good). (I am indebted here to questions asked by Jim Cargile.)

account should also preserve the rationality constraint whereby my desire to φ is somehow conditional upon φ-ing contributing to my real good. Can my rational desires be made out to be conditional in this way, without becoming unknowable in the process?

I believe that they can, if we make the conditionality on its contributing to my good part of the satisfaction-conditions of the desire, instead of making the having of the desire itself conditional upon its satisfaction contributing to my good. This can be effected by a kind of semantic ascent. The idea is that **R1** and **R2** are replaced by

R′ My desire is satisfied iff [A if $\mathring{A}$, $\sim A$ if $\sim \mathring{A}$],

which amounts to

R″ My desire is satisfied iff [A iff $\mathring{A}$].

That is, the satisfaction-conditions of my desire to φ, when this is subjected to the rationality constraint, is that A and $\mathring{A}$ have the same truth-value or, as a logician would say, A and $\mathring{A}$ are materially equivalent. Thus I can know what will satisfy my desire, even when the desire is subject to the rationality constraint. The satisfaction-conditions of my desire to φ, so characterized, are not conditional upon the inscrutable question of whether φ-ing contributes to my real good or not. But Socrates' conclusion is still vindicated, to the extent that if I φ, and φ-ing does not contribute to my real good, there is still something that would satisfy my desire, but I am not bringing about: the material equivalence of A and $\mathring{A}$. In this sense, *I am not doing what I want*.

That **R″** gives the satisfaction-conditions of my rational desire to φ and not its content, is clear from the following:

(i) it is absurd to claim that my desires are "really about" the truth-values of propositions;
(ii) the same formula gives the satisfaction-conditions of my desire *not* to φ, which cannot have the same content.

What's more, the satisfaction-conditions, so characterised, can have no explanatory force, since in that case the explanatory force of my desire to φ would be the same as that of my desire not to φ. How do we derive the content, and so something that can figure in an explanation of what I do, from the satisfaction-conditions? An obvious suggestion is to conjoin to **R″** a clause along the lines of 'it seems to me that $\mathring{A}$'. It is true that we thereby give up on the project of giving an account of an agent's real or rational desire, where that stands opposed to what seems good to the agent. But this is what we should expect: if we extrude all reference to

what seems good to the agent, we no longer have anything that can contribute to an explanation of what the agent does.

Notice that our formulation of the rational constraint on desires – that I envisage the satisfaction-condition of my desire that *A* as *A*'s having the same truth-value as *Å* – does not itself require that my desire that *A* be equivalent to, or necessarily conjoined with, a belief that *Å*, although some such belief will be invoked in any informative explanation of my rationally bringing *A* about.[21]

What is the upshot for our overall conception of the relation between desire and the good of our rejection of the Platonic conception of a "bare" desire for the good? This conception was rejected because it is a "wheel that turns independently of the rest of the mechanism": the bare desire for the good cannot figure in an explanation of anything we do, because it imposes on any such explanation a form according to which all of our rational desires are opaque to us. The explanation of action can only get started once we

(a) explain the conditionality of rational desire in a way which avoids appeal to a bare desire for the good and makes room for an agent's knowledge of what will satisfy his rational desires;
(b) add in facts about what end or action, within the agent's ken, *seems good* to the agent.

Rational action is inexplicable in the absence of appearances of the good. This suggests that, even once we have posited an objective good, a certain kind of concession must be made to the subjectivist: that wherever we appeal to this good in our accounts of rational action, it must be thought of as something which necessarily discloses itself to the agent *via possibly misleading appearances*. This in turn suggests certain asymmetries between theoretical and practical reason: the practical realm is partly shaped, as the theoretical is not, by an ineliminable provisionality and fallibility, and the only conception of the good available to practical reflection is *what these appearances are appearances of*. This is not to rule out the possibility of a direct intuition of the good to mystical consciousness, but it is to deny that such an intuition of the good could bring with it knowledge that could be put to use in practical reflection.

[21] The possibility of such an equivalence or necessary connection is denied by Lewis (1988), on the grounds that it conflicts with our best account of "probability kinematics"; the possibility is defended by Price (1989); see also Lewis (1996).

4. What Is Socrates Arguing Against?

Finally, I would like to turn to the question of what view Socrates is really opposing when he maintains that tyrants have no real power and don't do what they really want. In particular, I would like to warn against a temptation to see the issue between Socrates, on the one hand, and Polus and Callicles, on the other, as that of subjectivism *versus* objectivism about value. Since Polus and, by implication, Callicles have neglected the distinction between doing as one pleases and doing what, so far as possible, one can endorse as being really good, it may seem plausible to see them as anticipating the subjectivist conception of value as found in Hobbes, Hume and currently popular accounts of rational action in terms of the maximization of expected utility as defined by a subjective (i.e. agent-relative) utility function.[22] After all, Polus and Callicles, and even to an extent Gorgias himself, show by their admiration and envy of tyrants that they believe that the best life consists of doing what one wants. But this is crucially ambiguous as a conception of the best life. It may or may not involve a commitment to certain desires as being natural or best. If it does, the view is not really subjectivist after all, since it conceals a prior commitment to an objective standard of appropriateness for desires. If it does not, it is subjectivist and is committed to a formal, reconstructive account of the good, of the sort I began by sketching, in terms of the satisfaction of desires *whatever those desires happen to be for*.

It is pretty clear that the Gorgias-Polus-Callicles view is of the former kind, for (at least) two reasons. First, all three apologists for tyranny, to differing degrees, think of the tyrant as an objectively *grand* figure. That is, none of them sees the point of tyranny as enabling the tyrant to satisfy more of his desires independently of what those desires are for. On the contrary, they all have, again to differing degrees, a conception of what sort of desires a tyrant will naturally and appropriately have, and these involve the domination of others partly as an end in itself. Gorgias speaks of the orator appropriating the products of all other *technai*; Polus speaks with thinly-concealed envy of the crimes of the usurper Archilaus, even if his "false shame" prevents him from wholeheartedly endorsing injustice; and Callicles, of course, is explicit in his advocacy of the strong over the weak (even if he hasn't thought carefully about what this advocacy amounts to), and in the course of his argument with Socrates espouses a hedonism of a markedly objectivist kind. In all three cases,

[22] Penner sometimes gives this impression, e.g. (1991, pp. 148, 158).

the conception of tyranny as a good is rooted in a conception of value entirely independent of desire as such. The fact that a tyrant will, as a matter of fact, be able to satisfy more of his desires regardless of what they are for is therefore a red herring. Secondly, if any of them did hold a subjectivist conception of value, their response to Socrates' arguments about tyranny would be very different from what it is. Socrates' arguments obviously presuppose that there is such a thing as one's real, objective good – that is, the object of rational, unconditional, "real" want or desire. If Gorgias, Polus or Callicles were even flirting with a subjectivist conception of value, surely they would object to Socrates' presupposition as flagrantly begging the question against them. I suspect that the fact that they lodge no such protest does not show merely that none of them is a subjectivist. It shows that subjectivism is not even on the conceptual map. The idea that there is such a thing as one's objective good independent of whatever seems good at any particular time is a deep presupposition it would not occur to any of the speakers in the *Gorgias* – nor, I conjecture, to any of the ancients – to call into question. The subjectivist conception of value may well be another of the few philosophical positions that were not so much as formulated in antiquity.

Acknowledgements

I would like to thank Professor Myles Burnyeat for helpful comments on an earlier draft of this paper.

University of Bristol
Department of Philosophy
9 Woodland Road
Bristol BS8 1TB, UK
e-mail: J.A.Doyle@bristol.ac.uk

REFERENCES

Anscombe, G.E.M. (1963). *Intention* (2nd edition). Ithaca, New York: Cornell University Press.

Aristotle (1963). *Ethica Nicomachea*. Edited by: I Bywater. Oxford: Oxford University Press.

Arnauld, A. (1683). *Des Vraies et des Fausses Idées*. Cologne: Nicholas Schouten.

Austin, J.L. (1964). *Sense and Sensibilia*. New York: Oxford University Press.

Burnyeat, M.F. (1982). Idealism and Greek Philosophy: What Descartes Saw and Berkeley Missed. *Philosophical Review* **90**, 3-40.

Dodds (1959). *Plato: Gorgias. A Revised Text with Introduction and Commentary*. Oxford: Clarendon Press.

Doyle, J. (2004). Socrates and the Oracle. *Ancient Philosophy* **24**, 19-36.

Doyle, J. (2006a). The Fundamental Conflict in Plato's *Gorgias*. *Oxford Studies in Ancient Philosophy* **30**, 87-100.

Doyle, J. (2006b). On the First Eight Lines of Plato's *Gorgias*. *Classical Quarterly* (forthcoming).

Lewis, D. (1988). Desire as Belief. *Mind* **97** (387), 323-332.

Lewis, D. (1996). Desire as Belief II. *Mind* **105** (418), 303-313.

McDowell, J.H. (1994). The Content of Perceptual Experience. *Philosophical Quarterly* **44**, 190-205.

McTighe, K. (1984). Socrates on Desire for the Good and the Involuntariness of Wrongdoing: *Gorgias* 466a-468e. *Phronesis* **29**, 193-236. Reprinted in: Hugh H. Benson (ed.), *Essays on the Philosophy of Socrates*, pp. 263-297. Oxford: Oxford University Press, 1992.

Penner, T. (1991). Desire and Power in Socrates: The Argument of *Gorgias* 466A-468E that Orators and Tyrants Have No Power in the City. *Apeiron* **24** (3), 147-202.

Plato (1963). *Opera*, 5 vols. Edited by: J. Burnet. Oxford: Oxford University Press.

Price, H. (1989). Defending Desire-as-Belief. *Mind* **98** (389), 119-127.

Reid, T. (2002). *Essays on the Intellectual Powers of Man*. Edinburgh: Edinburgh University Press.

Segvic, H. (2000). No-one Errs Willingly: The Meaning of Socratic Intellectualism. *Oxford Studies in Ancient Philosophy* **19**, 1-45.

Tenenbaum, S. (1999). The Judgment of a Weak Will. *Philosophy and Phenomenological Research* **49** (4), 875-911.

Tenenbaum, S. (2007). *Appearences of the Good*. Cambridge: Cambridge University Press.

Weiss, R. (1985). Ignorance, Involuntariness and Innocence: A Reply to McTighe. *Phronesis* **30** (3), 314-322.

Weiss, R. (1992). Killing, Confiscating and Banishing at *Gorgias* 466-468. *Ancient Philosophy* **12**, 299-315.

Iakovos Vasiliou

VIRTUE AND ARGUMENT IN ARISTOTLE'S ETHICS

A definitive interpretation of Aristotle's account of incontinence would be an integral component in a satisfying and thorough understanding of his moral psychology. Accordingly, it has received considerable scholarly attention in recent years. While it is clear that Aristotle is critical of Socrates' denial of incontinence, and so is critical of his much-maligned idea that knowledge is sufficient for virtue, the precise nature and extent of his criticism has proven difficult to pin down, and scholars are still deeply divided on the proper interpretation.[1] My concern in this paper, however, is not with incontinence. I believe that the understandable attention that it has received has obscured another significant area of confrontation between Aristotle and Socrates, which has by contrast attracted little detailed study. I shall argue that Aristotle is centrally concerned with examining the role that knowledge plays in becoming and being good in the first place, and that his thinking on this topic can be shown to emerge in part from reflection on and reaction to Socratic ideas.

In addition to the significance that this topic has for understanding Aristotle's moral psychology, and for clarifying and deepening our understanding of the development of thought on these questions from Socrates to Aristotle, it will also bear on the issue of what Aristotle believes the role of *argument* is in becoming good. Insofar as knowledge is something conferred by argument and teaching, the magnitude of the role of argument in *becoming* good will wax or wane with the significance knowledge plays in *being* good. What Aristotle says about the place of argument in ethics can be illuminated by seeing in detail how it addresses questions and positions in the Socratic dialogues about the

[1] By 'Socrates' I am referring throughout to the character in Plato's dialogues, not the historical figure. It is a matter of some contention whether Aristotle's references to 'Socrates' refer to the historical or the Platonic figure; see Irwin (1995, c. 1), and Kahn (1996, pp. 79-87).

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 37-78. Amsterdam/New York, NY: Rodopi, 2007.

role of knowledge and argument in both becoming and being good. We shall see that Aristotle's engagement with Socrates in the *Nicomachean Ethics*,[2] is more extensive and pervasive, and to some degree more fundamental, than focus on the issue of incontinence alone would suggest. I shall begin by framing some familiar features of Socratic philosophy in a particular way.

1. Socrates on the Supreme Aim of Action and the Content of Virtue

The Socrates of Plato's dialogues is completely committed to the supremacy of virtue (SV) over all other aims one might have in action. One ought always, according to Socrates, to aim at virtue above all, and to (try to)[3] do the virtuous action rather than any other non-virtuous action, even if the virtuous action results in disgrace, impoverishment, or death. Socrates is especially clear about his total commitment to virtue in the *Apology*, although it is present, implicitly and explicitly, in many other dialogues as well.

Consider the following well-known passage:

> Perhaps someone might say: "Aren't you ashamed that you have pursued the sort of pursuit on account of which you are now likely to be put to death?" But I would reply to this with a just statement [δίκαιον λόγον] in saying, "You are not right, sir, if you think that a man who is worth even some little bit ought to take into account the risk of living or dying and not instead look to this alone when he acts: whether he is doing just or unjust things, the deeds of a good or a bad man." (28b5-9)[4]

Socrates' statement of SV here shows that SV is what I shall call an "aiming principle." An aiming principle tells the agent what *aim* she ought to have in acting; for example, to do the virtuous action. Socrates says that a man should "look alone" (μόνον σκοπεῖν) at whether he is acting virtuously or not. A person's *goal* ought to be to realize just and virtuous actions above all. SV does not by itself rule in or out any neutrally described action-type, and it leaves entirely up to the agent how to *determine* what the virtuous action actually is. So merely adhering to SV leaves open what sorts of considerations may be relevant in any

---

[2] Hereafter I will refer to this work as "*NE*." See references for ancient texts.

[3] Given Socrates' own ignorance (and that of all of those he has met) about what virtue is, one might after all fail to do what is in fact virtuous because of that ignorance.

[4] All references to Plato are by standard Stephanus page numbers: Burnet (1902-1906). All translations are my own, unless otherwise noted.

particular deliberation about what to do, as well as what particular actions such deliberation might yield. The pleasure or pain the action causes oneself or others, the financial cost, the risk one runs of life or death may all be relevant considerations in determining what the virtuous action actually is here and now. SV simply but importantly maintains that a person's supreme *aim* must always be to act virtuously, and not to save her life, or to cause pleasure, or to generate financial gain; it goes no distance towards helping an agent to determine what *is* virtuous or not in any particular situation. To have failed to achieve the aim of survival may appear to constitute the ultimately unsuccessful life. SV, however, denies this by claiming that acting virtuously trumps any other aims one might have in action, including even staying alive.[5]

Many readers come away from the *Apology* moved by Socrates' commitment to virtue. But surely a reflective person would then think that if virtue is of supreme importance, we ought to know what virtue is. The so-called "dialogues of definition" – in particular, the *Charmides*, *Euthyphro*, and *Laches* – attempt to tackle the question of *content*: Socrates addresses this topic primarily by means of his "what is F?" question, testing with the *elenchus* any person who thinks he has knowledge of virtue. A Socratic definition, briefly, will consist of a statement of the element that is common to all token actions of the type in question and that explains why the token actions are actions of that type. It is an assumption of Socrates' investigation that there is such a common element, and that, if someone knows what it is, he can put it into words.

In the *Euthyphro* (6d-e), for example, Socrates reprimands his interlocutor for offering an *example* of a (type of) pious action, rather than stating what piety is. Socrates wants to examine different actions in the world and be able to separate the pious ones from the impious by knowing what the distinctive and unique feature is that all and only pious actions have in common. In such cases Socrates is clearly trying to find a universal, what has been called a "Socratic form." So here is one type of question we can ask about virtue and the particular virtues: can we say what they are in the Socratic sense? Can we state a distinctive and unique feature that all courageous actions share, which makes them all courageous? This is one test for counting as knowing what courage is; I will refer to it as asking for a "Socratic definition," or simply a

[5] I argue for this interpretation and explain its significance in detail in Vasiliou (forthcoming, Chs. 1-2).

"definition."[6] Socrates apparently maintains that possession of such definitions is both necessary and sufficient for having "knowledge of virtue." The dialogues of definition fail to satisfy the criteria for a Socratic definition and so never arrive at an adequate account of virtue or of any particular virtue. We should recognize that the idea that knowledge of such an account is necessary for virtue is certainly contentious, albeit less so than the claim that it is sufficient.

One might, however, ask a different question. Rather than trying to offer a definition of courage in general, one might wonder what the courageous thing to do is in the here and now. This question is distinct, though clearly related to the issue of definition, for if one had a Socratic definition, then she could determine what the courageous action is here and now. Knowing a Socratic definition would be *one* way of answering a question about a token action in specific circumstances. But of course the question might be settled otherwise than by applying a general definition to particular circumstances. A person might, for example, have a magical intuitive grasp of what the virtuous thing to do is here and now, without appealing to any definition. A person might even simply guess, and get lucky. This issue arises for Socrates in the *Crito*, where he must answer, somehow, the question of whether leaving prison that very night is the virtuous course of action or not.[7]

In a short dialogue attributed to Plato, the *Cleitophon*, the eponymous interlocutor sharply distinguishes between two sets of tasks:

(1) turning a person towards virtue; that is, convincing a person that virtue is more important than anything else and so ought to be pursued above all;
(2) figuring out how to determine what virtue is.[8]

According to Cleitophon (407a7), Socrates does an excellent job at persuading people to pursue virtue and the care of their souls, but is utterly unhelpful when it comes to saying what virtue actually is. This dialogue highlights the distinction we have seen between taking virtue as one's supreme aim, and determining, whether in some particular case or

[6] It should be clear that "definition" is simply shorthand for a reading of the *Euthyphro* passage. We need not explore the issue of Socratic "definition" further here. For one recent discussion of these questions, and additional secondary literature, see Silverman (2002).

[7] While it is disputed whether the *Crito* presents his real reasons for remaining in jail, Socrates *does* manage to arrive at a decision (and so arrives at an answer to a practical question). See Vasiliou (forthcoming, Ch. 2).

[8] Slings (1999) is a recent study of the dialogue, which includes a discussion of its authenticity.

in general, what virtue is. Because Socrates is so useless with this substantive question, Cleitophon is forced to conclude that either Socrates' ability to champion justice does not in any way imply that he knows what justice is, or else Socrates is simply unwilling to tell him (410c). Cleitophon, frustrated and disappointed by Socrates' failure in this regard, ends by threatening to become a pupil of Thrasymachus instead, since at least he provides an answer to this question.

I think that the distinction highlighted in the *Cleitophon* is fundamental for Aristotle as well, since he distinguishes sharply between questions about what eudaimonia is and questions about what virtue is. He supplies arguments about the first, but simply describes characteristics of the second. Moreover, we shall see that his account of habituation, central to his discussion of virtue of character, is developed explicitly in opposition to Socrates' search for knowledge of what virtue is. While he follows Socrates in being committed to a life of virtue, he disagrees with Socrates' account of how to determine what virtue and virtuous actions are. We shall see this distinction reflected in what can seem to be Aristotle's conflicted stance regarding the role and purpose of argument in ethics. I shall ask two questions about argument:

(1) when and where does something recognizable as argument occur in the *NE*, and when is it conspicuously absent?
(2) what is the role of argument in being and becoming good?

2. Putting *Logos* Down: The Limitations of Argument for Ethics

In the *NE*, Aristotle says more than once that the study of ethics is about actually becoming good and acting well, not simply about knowing the good (1095a5-6, 1103b27-8, 1179a35-b3)[9] (see also Aristotle *Eudemian Ethics*, 1216b21-5, discussed below). Aristotle also makes deflationary remarks about the ability of arguments to make a person a good (II.4, X.9), and, he requires the listener of his lectures to be well brought up before attending them (1094b28-1095a6, 1095b3-8, 1179b25*ff*).[10] It is reasonable to connect these thoughts. Arguments and teaching can supply

[9] All references to Aristotle are by standard Bekker page numbers: Bekker (1831-1870).

[10] Burnyeat (1980) offers an interpretation of Aristotelian moral development arguing against the idea that Aristotle is trying to convince a sceptic of the correctness of his own substantive conception of eudaimonia. In Vasiliou (1996), I attempt to explain *why* Aristotle believes that arguing with those who are not well brought up does not work. McDowell (1996a), discussed in §6, provides an account of the role of upbringing in the formation and development of practical wisdom.

a person with knowledge, but, in the end, ethics is not (primarily) about knowing, it is about acting. Thus arguments in ethics do not have the same value that they have in other disciplines, like natural science or metaphysics, where the goal is explicitly the acquisition of knowledge.

If arguments do not make us good, then what does? Aristotle's well-known answer is habituation. In II.1, we learn that habituation is the process of acquiring "habits" (ἕξεις) by repeatedly engaging in actions of a similar type.[11] The type of habituated state a person acquires is determined by the type of actions she engages in: for example, one becomes brave by engaging in brave actions. Habituated states of character are not states we are born with. Rather, we are born with the *ability to acquire* such states, states that form a person's "second nature" and that are difficult, though perhaps not impossible, to change.[12] The more frequently a person does a certain type of action, the more one becomes that sort of person; and then, in turn, the more one becomes that type of person, the more one is liable to do that sort of action. In other words, momentum is a significant force in the initial formation or subsequent alteration of character (II.2, 1104a33-b3). Although it may be quite difficult to begin to acquire a new habit, it will, if one continues on the same course, get easier and easier, and correspondingly more and more difficult to act contrary to the newly acquired habit (*Categories* 13a23-31). For this reason Aristotle emphasizes that it is most important to acquire the right sort of habits right from youth (1103b23-5, 1104b11-3). In addition, most scholars agree that despite how it might look after a reading of II.1 on its own, Aristotle's account of habituation is not simply the mindless training of one's appetites – training that, for example, might apply equally well to a dog. A person's habituation does not simply leave him with desires and motivations to do certain sorts of actions and avoid others, but it provides him also with some sort of cognitive content or ability, although commentators vary greatly about just what this amounts to.

Aristotle also contrasts coming to be good by habit with coming to be good by teaching and by nature. He associates argument (*logos*) with teaching, saying that argument and teaching do not work in all cases (X.9, 1179b20*ff*). After completing his account of habituation in II.4, and vindicating his claim from II.1 that a person becomes just, for example,

---

[11] The "similarity" in type does not necessarily refer to any physical or psychological similarity among different token actions. The similarity will be in terms of the common description of them as just, brave, kind, or whatever. See Broadie (1991, p. 108).

[12] See Di Muzio (2000) who argues persuasively that Aristotle even allows that a vicious person is not entirely incurable.

by doing just actions, he warns against those who would try to substitute this "doing" or practicing with a specious kind of philosophizing (1105b12-8). He says that instead of doing the appropriate actions over and over, which will actually make a person good, most people "flee for refuge into *logos*," and behave like sick people who listen to their doctors, but then follow none of their directions. I shall consider this passage below in greater detail, but the overall point of all of these passages is not difficult to grasp. Becoming good, the goal and task of ethics, takes hard work and practice. Ethics is about action, not knowledge. Most people, however, prefer to discuss what being virtuous is, and falsely think that they are thereby making themselves good and engaging in philosophy.

I have tried to establish quickly a connection between Aristotle's idea that ethics is about action, not knowledge, and his idea that people are made good not by argument and teaching, but by habituation. Why does Aristotle put as much emphasis on these points as he does? For one thing, he is sensitive to the different methods, goals, and degrees of precision required by different subjects. To misunderstand these issues is a sign of lack of education (ἀπαιδευσία; see, e.g., *Met*. IV.4, 1006a6-7). At the beginnings of the *Metaphysics*, *Physics*, and *De Anima*, he explicitly states that the goal of such works is to achieve knowledge. So when he highlights, by contrast, the idea that knowledge, something conferred by argument and teaching, is *not* the goal in ethics and politics, he is in part simply following his own advice about being careful to mark the different goals appropriate to different disciplines. But I shall show that Aristotle also has a specific target in mind in his emphasis on the essential role of habituation in becoming good, and in his associated downplaying of the value of argument: Socrates.

3. Aristotle on Socrates on What Virtue Is: The *Eudemian Ethics* and *Magna Moralia*[13]

Aristotle's account of habituation may be understood to be motivated by a criticism of Socrates on the question of what virtue is. Not only is his account of habituation plausibly understood as developing in opposition to Socrates, but so are his criticisms of the role argument may play in

[13] Hereafter I will refer to these works as *EE* and *MM* respectively.

becoming good. In this section I shall focus on comments and criticisms of Socrates in the *EE* and *MM*.[14] The *EE* contains a central text (EE1):

> Socrates the elder[15] thought that the end [of life] was to know (τὸ γινώσκειν) virtue, and he used to inquire what justice is, and what is courage, and [what is] each of its [virtue's] parts. It is reasonable that he made such inquiries [ἐποίει γὰρ ταῦτ' εὐλόγως], for he thought that all of the virtues were branches of knowledge [ἐπιστήμας], so that both knowing justice and being just would coincide at the same time. For as soon as we have learned geometry and construction, we are builders and geometricians. For this reason he sought what virtue is, but not how it arises and from what sources. This [the way Socrates proceeded] is fitting in the case of the theoretical sciences, for there is nothing else that belongs to astronomy or knowledge of nature or geometry except knowing and studying [θεωρῆσαι] the nature of the objects which fall under the sciences [ἐπιστήμαις] – not that anything prevents them from being coincidentally useful to us for many of the things we need. In the case of the productive sciences, however, the end is different [ἕτερον] from the science [ἐπιστήμης] and knowing [γνώσεως]: for example, health [is the end of] medicine, and good social order [εὐνομία], or some other different thing, is [the end of] politics. Now it is fine to know each of the fine things; nevertheless, concerning virtue, at any rate, it is not knowing *what it is* that is most valued [τιμιώτατον], but knowing *from what sources* [ἐκ τίνων] it arises [ἐστίν]. For we do not wish to know what courage is, but to be courageous, nor [to know] what justice is, but to be just, just as also [we wish] to be healthy rather than to know what being healthy is, and to be in a good condition rather than to know what it is to be in a good condition. (1216b2-25)

I shall note three points right away. First, the end of this passage explicitly connects criticism of Socrates with the point that "ethics is about action, not knowledge," familiar to us from the *NE*. Second, and more importantly, Aristotle concludes by devaluing the Socratic question of "what is virtue?" and replacing it with his own question, "how does virtue arise?". The "most valued" knowledge is knowledge of how virtue arises. Aristotle's answer to this is, of course, habituation. So, although there is no mention of Socrates by name in the relevant parts of the *NE* (II.1-4), it would be reasonable to read the account there as part of an explicit reaction to his thinking. Finally, the passage suggests that

---

[14] Sherman (1989, c. 5), begins her discussion of Aristotelian habituation with the quote below from the *EE*, and uses the Socratic position as a contrast to Aristotle's (pp. 157, 161, 199). But she does not discuss the comparison between Aristotle and Socrates in any detail.

[15] The fifth century philosopher who is used as a character by Plato.

Aristotle's view about the relation of virtue to knowledge will be complex. He neither dismisses knowledge as irrelevant to virtue, nor does he embrace it as the key. Certain types of knowledge – knowledge about how virtue is arises – are very important to being virtuous, but others – perhaps Socratic definitions – less so. The simplistic slogans of knowledge as necessary or as sufficient for virtue will have to be expressed more precisely if we want to capture accurately Socrates' and Aristotle's positions and the relationship between them.

In commenting on this passage, Michael Woods points out that Socrates can agree with Aristotle that the goal of ethics is to be virtuous, not simply to know what virtue is (Woods 1982, p. 61). Since this seems rather obvious, there is a question about why Aristotle appears to criticize Socrates on this front, and to present the point as though it were a rejection of his position. According to Woods, what explains why Aristotle makes this point is Socrates' belief that knowledge is *sufficient* for virtue, most famously attributed to him by Aristotle in *NE* VII.2 (1145b23-7), a passage from the "common books." Woods is somewhat puzzled by the passage's apparent criticism of Socrates' "what is F?" question, pointing out that Aristotle himself arrives at an account of how virtue is achieved through discussion of what virtue is:

> Hence, nothing in the argument of the present passage tends to show that the method to be employed is totally different in kind from that directed towards Socratic definitions. At 1216b20-21, Socrates is criticized for asking what virtue is, instead of trying to see how it is achieved. But Aristotle's own answer to the latter question is arrived at, in II, by way of an investigation into the question what virtue is. Throughout the *EE* and the *EN* Aristotle poses, and attempts to answer, questions of the "what is X?" form. It is difficult to see how ethics can contribute to the practical aim which is here insisted on except by answering theoretical questions of this kind. (Woods 1982, pp. 61-62)

Since Woods believes that Aristotle proceeds to supply his own answer to the question "what is virtue?," and also that ethics cannot contribute to a practical aim without this sort of investigation, he interprets the passage simply as a rejection of the sufficiency claim and then gently criticizes Aristotle for pushing this reasonable point too far by appearing to suggest that the entire investigation into what virtue is is not particularly important, or perhaps could be supplanted.

The problem with the passage as it stands, however, is that at 1216b20-5 Aristotle clearly states that determining what virtue is is less worthwhile than determining how virtue comes about. He exacerbates this by presenting a contrast between knowing what justice is and being

just and knowing what health is versus being healthy. This analogy is particularly troublesome. A fortunate person might live an entirely healthy life, without ever inquiring into or learning about what health is. Knowledge of what health is would be for such a healthy person not even *necessary* for his health, as long as a person knows what the sources of health are – or exhibited that knowledge in her actions.

In one respect Woods is clearly correct: Aristotle does go on to ask and then answer questions precisely of the form, "what is virtue?".[16] While Aristotle's investigation of this question might not be "totally different in kind" from Socrates', we shall see below that it differs in significant ways. If so, there might be more to EE1 than simply a rather badly expressed criticism of the sufficiency claim. Indeed I find no evidence that EE1 aims at criticizing Socrates' view that knowledge is sufficient for virtue. Of course, Aristotle *does* criticize this view elsewhere, but if we read it into EE1 we shall miss a distinct and substantial criticism of Socrates' method of investigation into virtue, which will play an important role in understanding Aristotle's own attitude towards virtue. According to Aristotle, the Socratic search for an answer to the "what is F?" question will turn out to be wrong both practically speaking (insofar as one wants to become and make others good) and theoretically speaking (insofar as one wants to know what virtue is and how it arises).

In the *MM*, indeed, we find that only *part* of the criticism of Socrates is aimed at his belief that knowledge is sufficient for virtue, and therefore that incontinence is impossible. The author[17] attacks Socrates for making the virtues all branches of knowledge, and for therefore locating all of the virtues in the rational part of the soul. MM1:

> For [Socrates] made the virtues branches of knowledge [ἐπιστήμας]; but it is not possible that this is so. For all of the branches of knowledge are with *logos*, but *logos* arises in the intellectual part of the soul. Therefore all of the virtues, according to him, [arise] in the reasoning part of the soul. Therefore it follows for him that by making the virtues branches of knowledge, he does away with the non-rational part of the soul, and by doing this, he does away with both emotion and character. (1182a17-23)

Mention of a non-rational part of the soul of course triggers thought of the problem of incontinence. Once a thinker acknowledges "parts" of the soul, as Plato does in *Republic* IV or in the *Phaedrus*, the possibility

[16] For example, *EE* II.1, 1220a13*ff*, *NE* II.5, 1105b19*ff*. I shall discuss how Aristotle answers this question below, §5.

[17] I will use "the author" with respect to *MM*, since most scholars do not believe that the text is actually by Aristotle. The question of authorship is not important for my argument.

of conflict, and so of incontinence, arises. But we should note that the author's criticism of Socrates here is not explicitly about the sufficiency claim or about incontinence. And even if we were right to see it being hinted at, the criticism of Socrates in the passage is certainly not *limited* to this subject, but extends more widely. The final sentence complains that Socrates' utter neglect of the non-rational part of the soul leads not to a failure to account adequately for incontinence, but to a neglect of emotion (πάθος) and character (ἦθος). As we have seen, habituation is the formation and development of character. Although there is no elaboration of the point, MM1 says that Socrates' focus on the virtues as branches of knowledge leads to his missing the significance of character. This passage should be read, then, as criticizing Socrates for missing the relevance and importance of habituation because of his belief that all of the virtues are branches of knowledge. As in EE1, this point is distinct from any about incontinence.

Consider a second passage from the *MM* (1183b9*ff*) (MM2):

> Nor was Socrates correct when he made the virtues branches of knowledge [ἐπιστήμας]. For he thought that nothing should be in vain [μάτην], but from the virtues being branches of knowledge, it follows for him that the virtues are in vain. Why is this? Because in the case of branches of knowledge [ἐπιστήμαι] it follows that at the same time a person knows what the knowledge is, he is also a knower – for if someone knows what medicine is, this man is also right away a doctor, and the same with the other branches of knowledge. But this does not follow in the case of the virtues. For it is not the case that if someone knows what justice is, he will straightaway be just, and the same with the other [virtues]. Therefore it follows both that the virtues will be in vain and that they are not knowledge. (1183b9-17)

It is clear that here the author *is* attacking the idea that knowledge of what justice is is sufficient for being just. The argument seems to be that Socrates, like everyone else, believes that possession of the virtues is sufficient to yield virtuous action; it makes no sense to say that a person is just and yet does not act justly. So if, as Socrates believes, the virtues are simply knowledge, then knowledge alone must make a person virtuous. But the author takes it as obvious that knowledge of what virtue is does *not* by itself make one virtuous, and so Socratic virtues will not, after all, *be* virtues in the sense just described, but will be "in vain" and useless.

Unlike in MM1, MM2 is explicitly concerned with the sufficiency claim. It may at first seem straightforwardly similar to EE1, which might lead one to conclude that EE1 is also discussing sufficiency, insofar as

they both discuss purported cases in which knowing something *ipso facto* makes one a certain kind of person. But this would be to fail to notice the striking contrast between the two passages. Whereas MM2 moves to attack Socrates for his sufficiency claim, EE1 moves in a different direction and uses slightly different examples. In EE1 the division is made between theoretical sciences (which are all about knowing and studying) and productive sciences. The distinction Aristotle intends is difficult to pin down precisely (see Woods 1982, p. 62, for some discussion). But we should note that when Aristotle uses the example of medicine in EE1, he looks at it as a productive science whose "different" product is health. In MM2 the issue is not the "product" of medicine, but the point that a knower of medicine is *ipso facto* a doctor. EE1 does not discuss the relationship between knowledge of medicine and being a doctor, but is concerned only with the product of medicine: generating health. So MM2 is instructive insofar as its concern *is* with the sufficiency claim, since it enables us to see how it treats the similar examples differently from EE1.

A third passage from the *MM*, parallels the EE1 most closely (MM3):

> Thus Socrates did not speak correctly when he declared that virtue is reason [λόγος], for [Socrates thought that] it is of no benefit to do courageous and just things, when one does not know and does not choose with reason. Thus he said that virtue is reason – incorrectly. Those who [speak about these things] now, [speak] better. For they say that this is virtue: doing fine things [τὰ καλά] according to correct reason [κατὰ τὸν ὀρθὸν λόγον]. But not even these people speak correctly. For someone might do just things without any choice [προαιρέσει μὲν οὐδεμιᾷ] nor with knowledge [γνώσει] of fine things, but with some sort of non-rational impulse [ὁρμῇ τινι ἀλόγῳ], and [do] these things correctly and according to correct reason (I mean that he acted in the way that correct reason would command). But nevertheless such an action is not praiseworthy. But it is better [to say], as we define it, that [virtue] is the impulse towards the fine with reason, for such a thing is both virtue and praiseworthy. (1198a10-23)[18]

[18] The passage is striking for its claim that Socrates believes that virtue is *logos* (but see below, n. 22). This differs from the more typical formulation: that virtue is knowledge (ἐπιστήμη; as in EE1, and in MM1 and MM2). It appears, however, that the difference is not too significant. If we think of the *logos* referred to as the account of what virtue is, then for Socrates virtue consists in the possession of that account, since possession of such an account would give the possessor knowledge. When one decides "with *logos*," then, he decides according to the right account. This fits with the author's criticism of the "men of today" who equate virtue with acting according to right reason. The objection to this is that it is ambiguous insofar as it leaves open a possibility that is too anti-Socratic:

According to the author, Socrates is wrong to dismiss the simple doing of courageous and just things as worthless simply because the person "does not know and does not choose with *logos*." The text claims that it is *because* Socrates believed that doing what is just or courageous without the accompanying knowledge is worthless that he also then equated virtue with *logos*: "*For this reason* he said that virtue is *logos* – incorrectly." If one believes that there is no value to doing a "virtuous" action without knowledge and "an account," then one could see that virtue might be thought to be the same as that knowledge. The author does not say here what he takes the value of such action to be, but we might naturally think of habituation. On the *NE* account, at least, we acquire (at least some) knowledge and the proper decision (προαίρεσις) by doing what the virtuous person would do over and over, without yet possessing the knowledge or decision ourselves. We can understand MM3 as making the same criticism of Socrates as we see in EE1. In MM3 the question of the sufficiency of knowledge is not mentioned; but there is a distinct criticism of Socrates' account of virtue. Since Socrates neglects the question of how virtue arises, he fails to notice the value of doing virtuous actions even if one does not possess knowledge of virtue and so does not possess correct decision. Socrates' failure is even more perverse, by Aristotelian lights, since it is the very doing of these actions that that is necessary for yielding the proper decision and the knowledge which Socrates sought (see §4).

What can we conclude from these texts? I have tried to show that Aristotle's criticism of Socrates' search for an answer to "what is virtue?" is not limited to the point that knowledge of the answer would fail to be sufficient for virtue. Aristotle has a more extensive criticism as well. Socrates' focus on knowledge of what virtue is leads him to miss a more fundamental and important question about how virtue arises. This failure, in turn, leads Socrates to ignore the importance of character. I am suggesting that Aristotle is not saying simply that Socrates' inquiry was good and valuable, only he did not go far enough, or needed something additional. The passages from the *EE* and *MM* suggest that there is something more deeply wrong with Socrates' project according to

---

one could be virtuous simply by doing what the virtuous person would do even if one does it "by nonrational impulse." The author then, in typical Aristotelian fashion, reconciles the two extreme positions – the first which overvalued the significance of the possession of *logos* (and made virtue identical to it), and the second which made action in accordance with right reason sufficient for virtue (thereby rendering possession of *logos* unnecessary to virtue). So the *MM* position is that true virtue requires a person to be motivated to virtue while in possession of right reason; thus the "knowing" aspect is necessary to virtue.

Aristotle. It is true that, as Woods notes, Aristotle himself *does* explicitly inquire into "what virtue is." But we shall see that the *NE* will devalue the Socratic search, and that Aristotle's own answer to what virtue is will turn out to be quite un-Socratic.

4. Habituation as a Replacement for Socrates' "what is F?" Question: *NE* II.1 and II.4

What happens to Socrates' "what is F?" question in Aristotle? It is clear, at least, that Aristotle is criticizing Socrates in EE1 for his overemphasis on Socratic definitions. He wants to supplant or replace the Socratic "what is F?" question with a question about the sources of virtue – the "most worthwhile" question. As I said above, Aristotle seems to disparage that knowledge as not only insufficient but also as relatively unimportant.[19] In thinking about Aristotle's own question about how virtue arises, and about his own answer to the question of what virtue is, we should keep in mind Socrates' (and his interlocutors') failure ever to provide any Socratic definitions. Such definitions would supply universal principles that could be applied to determine whether a particular action is virtuous or not. Aristotle will not provide the definitions that Socrates sought. When Aristotle asks his "what is virtue?" question, we shall see that it is quite different from Socrates' and brings quite different consequences. We have already briefly surveyed Aristotle's account of habituation; I shall here examine the structure and argument of II.4 in some detail.

II.4 poses a challenge to Aristotle's account of the acquisition of virtue from II.1, where he draws an analogy between the acquisition of a techne and the acquisition of virtue: both are acquired through repetition of like actions.[20] How can repetition of like actions give rise to a

[19] Am I saying that Aristotle denies that knowledge is necessary for virtue? Depending on how it is spelled out, this can be a very weak claim. Even if Aristotle acknowledges that knowledge is necessary for virtue, he might well think that the knowledge that is necessary is easily attainable, something that most people would know (see discussion below). That is what I mean by calling it "unimportant": it is not that one could do without it, it is simply easily available. I think this is suggested by Aristotle's use of the word 'τιμιώτατον'. Since the knowledge in question is so available (although necessary) it is really not where the difference between the virtuous person and others comes up.

[20] I shall simply write 'techne' for τέχνη (technai, plural) without long marks; typical translations are 'art', 'skill', or 'craft'. To acquire a techne is to acquire a body of knowledge. The focus in II.1 and II.4 on the similarities and differences between virtue

particular state, if possession of that state is necessary before one can perform such actions? Aristotle replies that in fact prior possession of the state is *not* necessary for such actions, "not even in the case of the technai" (1105a21-2). One can do the proper action in a craft either by luck, or by following the instructions of someone else, and the same is true for virtuous actions. Aristotle insists that this does not make one a craftsman. To *be* a craftsman one must not only do what the craftsman would do, but also do it *in the way that the craftsman would*, which consists in doing it as an expression of knowledge that is in him (1105a23-6). Thus far in II.4 Aristotle is expanding on the *similarity* between virtue and techne from II.1. Habituation is possible because a person can do the right action (either in the case of craft or in the case of virtue) without himself being a craftsman or a virtuous person, but only because he is either lucky or because he follows the instructions of someone else, who, presumably, does know what the proper actions are. Anyone who "does the correct action" in this sense, however, is *neither* a craftsman *nor* a virtuous person. Even in the case of the crafts one must do the action "as the craftsman would" – that is, do it as a result of knowledge that is in one (not simply by luck or by following another's directions) – in order to be a true craftsman.

II.4 allows then that even the vicious person may do a virtuous action in a minimal sense, either by accident or because he follows the instructions of someone else, by doing the action, described in non-moral terms, that a virtuous person would do in some circumstance, for example, putting out a fire, or giving ten dollars to someone. This is a critical aspect of Aristotle's account of the sources of virtue. The fact that a right action can be communicated to a learner in non-moral terms, and that it can be done entirely independently of the state of the agent, is important for the possibility of moral education. A child who has as yet no grasp of the concept of being noble can be told that a good boy here and now shares his cookie. It may be true that the concept of a "good boy" or of "kindness" cannot be *reduced to* any purely non-evaluative account, an account that does not include any ethical or moral terms; it is not always an act of kindness to share what one has. But that it is possible in a concrete circumstance for the virtuous action to be describable in such terms – here and now the kind act is to give away half of one's cookie – in part makes moral education and the acquisition of

---

and techne provides additional evidence that Aristotle is reacting to Socrates, who analogizes virtue to techne in the course of trying to find knowledge of virtue.

the virtues possible. It is also what makes character change later in life possible, even if unlikely.[21]

Only *after* discussing this additional similarity between virtue and craft does II.4 turn to the *disanalogy* between craft and virtue: for craft, the excellence of the product consists solely in the product, while the excellence of the virtuous action is also, partly, a function of the state of the agent. While it is sufficient for a skill to be executed excellently if its product is excellent, it is not sufficient for an act to have been done virtuously for a person simply to have done what the virtuous person would do. A shoe might, by accident, be an excellent shoe (we can determine this by examining the shoe); but a virtuous action cannot be virtuous by accident. We should remember, however, that even though an accidentally made excellent shoe is fully excellent, it has nevertheless not been made *as the craftsman would*. The claim is only that one does not need to execute a craft as a craftsman would in order to get an excellent product. One does, however, need to execute the craft as a craftsman would to be oneself a craftsman. Aristotle then explains that for an action to be virtuous one must not only *do* what the virtuous person would do, but do it *as* the virtuous person would.[22] This famously involves three conditions on the agent (1105a31-33):

---

[21] See DiMuzio (2000). Contrast Sherman (1997, p. 81): "But profound character change is not something Aristotle ever really envisions as a possible moment of adult life. It is certainly not something that philosophy as argument can undertake." I shall consider the second sentence below. We shall see too (§§6 and 8) that McDowell (1996a) believes that Aristotle's account of habituation renders the content of one's moral outlook "fixed." This will be a central point of contention. I shall argue that while habituation determines the content of virtue for Aristotle, he believes that it is argument that "fixes" it.

[22] Whiting (2002c, p. 276), writes: "[. . .] someone who chooses to perform an apparently virtuous action on account of the utility or pleasure such an action provides for her either fails to perform a genuinely virtuous action or performs such an action only coincidentally." I agree entirely with this: the motivation's being of the right sort is an essential feature of the action's belonging to the type "virtuous." What Aristotle needs, however, to solve to the puzzle of II.4 is not only a separation between virtuous action and motive but also the ability to describe the virtuous action, the action to be done, without using ethical terms. This is what makes it possible for the virtuous and non-virtuous agent to, in one ordinary sense of the expression, do the same action – for example, to share half their sandwich, even though it will only be a truly virtuous action if the agent is motivated in the appropriate way. I think this helps to alleviate a potential problem with circularity that Whiting sees for Aristotle's account (p. 277): "[. . .] he seems to want to define virtuous actions in terms of the virtuous agent (as the sorts of actions a virtuous agent would perform) and to define the virtuous agent in turn in terms of virtuous actions (as the sort of person who routinely performs virtuous actions for themselves)." I do not think that Aristotle tries to define virtuous actions in this way if "define" means to state their essence – in Socrates' terms, to say "what virtue is." What the virtuous agent does is

(1) he acts knowingly;
(2) he has decided on it and decided on it for its own sake;
(3) he acts from a firm and unchanging character.

I want to focus on (1), which is usually passed over without comment in the literature.[23] By itself this apparently innocent idea – that the virtuous person must act knowingly – turns out to be rather puzzling. Aristotle's subsequent elaboration of it, adding yet another point of comparison with technai, seems to add to the problems:

> On the one hand these [three conditions] do not count for having the *crafts*, except for the knowing itself; but for having the *virtues*, the knowing has little or no weight[24] [οὐδὲν ἢ μικρὸν ἰσχύει], but the other [two conditions] count [δύναται] not for a little but for everything [οὐ μικρὸν ἄλλα τὸ πᾶν], [the two conditions] which arise from doing just and temperate things many times. (1105a33-b5)

However we finally understand this passage in detail, it is clear that Aristotle is downplaying the significance of knowledge for the case of virtue, which ought to remind us of the criticism of Socrates, especially in the *EE*. The passage also brings in a second contrast between virtue and the crafts: while the latter two conditions do not count for crafts, "except for the knowing itself," for virtuous action the second two conditions are "everything" and the first counts for "little or nothing." Since this is the final point of comparison between virtue and craft, I shall summarize the similarities and differences:

(S1) Both are acquired by repetition of similar activities (from II.1).
(S2) One can do what the virtuous person or what the craftsman would do, without himself being either virtuous or a craftsman (namely by luck or by following someone's instructions). In neither case is one acting *as* the craftsman or the virtuous person acts (II.4).

(D1) The excellence of the product of a craft is intrinsic to the product, and has nothing to do with the state of the producer. By contrast, the excellence of an action (the "product" of the virtuous person)

---

virtuous by definition, but not *because* she does it; the explanation is the other way round, for Aristotle believes that virtue is objective. The virtuous agent does what she does *because it is virtuous*. Thus when Aristotle says that virtuous actions are those the virtuous agent would perform, that is simply a shorthand description, not an attempt at providing a definition. So on my view there is no circularity here but there is an outstanding problem: how does the agent determine what the virtuous action *is*?

[23] Exceptions are Broadie (1991, pp. 85-86), and Williams (1995, pp. 14-15).

[24] Irwin (1999) has: "the knowing counts for nothing or [rather] for only a little."

does not contain its excellence solely in its intrinsic features; its excellence is also partly determined by *how* it was done (II.4).

(D2) To do something as a craftsman would do it, "knowledge itself" counts but decision and character do not. To do something as the virtuous person would, the knowledge counts "little or nothing" while the decision and stability of character count "for everything" (II.4).

What is the "knowing" that counts little or nothing? We might think that it is simply the idea that one must know what one is doing; that is, meet some sort of ordinary conditions for intentionality. The point would be that I could do what the virtuous person would do, but without being aware that I am doing what the virtuous person would do. I might save someone's life by pushing her out of the way of an oncoming bus, without intending to save her life; perhaps I simply pushed her so I could be first in line for pizza. Sarah Broadie brings up the example of Oedipus' ignorance that the man he is killing is his father, and Bernard Williams offers the case of someone who sends a donation to a hospital by accident, as a result of putting a check in the wrong envelope (Broadie 1991, p. 85; Williams 1995, p. 14).

Although Williams thinks that this is the sort of knowledge Aristotle refers to, he objects that its relevance is "to an earlier question: not whether the V[irtuous] act was done as a V person would do it, but whether an even minimally V act was done at all. [. . .] These matters of intention are importantly different from questions of motive." While Williams is surely right that questions of intention are different from questions about motive, Aristotle uses neither of these terms. He is interested in the distinction between the virtuous *action* (conceived of, as we have seen, on analogy with the product of craft) and the *condition* (πῶς ἔχων, 1105a31) of the virtuous agent. I see no reason why Aristotle might not include in the description of the agent's condition features of *both* his intentional and his motivational states, distinct though these may be. Williams assumes that an entirely unintentional action cannot be "even minimally virtuous." But if all that is meant by "minimally virtuous" is doing what the virtuous agent would do as opposed to doing it *as* the virtuous person would do it (which is the contrast Aristotle is drawing), then one can of course *do* what the virtuous agent would do, without even being aware of it. Williams should not equate "doing an action as the virtuous person would do it" with the issue of motivation alone.

The correctness of Williams' objection aside, he concludes that the "knowledge mentioned in [the first] condition is everyday knowledge

relevant to effective intentions, as it is with technai" (p. 15). I shall call this the "awareness interpretation"; in short, the virtuous agent must be aware of what he is doing and not, for example, sending someone a check without realizing it. So far I have attempted to rescue this interpretation, which Williams endorses, from the accusation that it fails to be relevant to the question of "doing the virtuous action as the virtuous person would do it." On my reading, "what the virtuous person would do" is simply the action as described in non-ethical terms: return the money, remain with an injured person, and so on.

I do not believe, however, that the awareness interpretation can be the correct understanding of the knowing condition, despite its initial plausibility. A problem is that it makes little sense of D2. For the knowing to be "what counts" in the case of crafts, it surely isn't simply the knowledge that what I am doing is building a house; it must be the expert knowledge of *how* to build a house, not simply the "everyday knowledge" of what a person is doing. Further, if it *were* simply the "awareness" of what one is doing – building a boat, standing firm in a dangerous situation – then such awareness would seem to be equally important (and/or unimportant) for both crafts and virtue. Broadie suggests that the contrast consists in the fact that the craftsman has a more detailed and thorough knowledge about the action in question than the layman, but it is not a matter of the knowledge of the virtuous person that makes him particularly superior to the non-virtuous person (Broadie 1991, p. 86). Unfortunately understanding the knowing as a contrast between "expert knowledge" and ordinary knowledge runs into trouble as well. If we think of the corresponding "expert knowledge" of the virtuous person – namely moral knowledge – then *it*, surely, cannot count for "either little or nothing." The special knowledge of the person of practical wisdom (ὁ φρόνιμος), however we understand it, will, as Williams says, "make *all* the difference" (Williams 1995, p. 15).

I think therefore that the "knowing" referred to here must be knowing in some weaker sense than expert knowledge, but not so weak as to be simply awareness of what one is doing. It involves correctly identifying what ought to be done. At VII.1 (1145b12-3) Aristotle calls the continent and incontinent types "knowing" (εἰδώς), along with the virtuous person. In some sense all three of these character types have knowledge about what action ought to be done.[25] If this sort of knowing is meant in II.4 as

[25] I say "in some sense" since there is disagreement among McDowell (1996b) and (1998), Broadie (1991, c. 5), and others about whether the "cognitive" state of the virtuous person exactly matches that of the continent or incontinent person, despite the fact that they are

well (both passages use εἰδώς), then we might understand more easily Aristotle's claim that the knowledge by itself might be worth little or nothing since many types of people possess this knowledge, but fall short of being good.

This interpretation coheres nicely with the critical note on which Aristotle ends the chapter. Before he turns in II.5 to the question of "what virtue is" – the Socratic-sounding question that in the *EE* he disparages – he offers one final criticism aimed at "the many."

> The many, however, do not do these things [engage in the repeated actions that would actually make them good], but instead they think that they are doing philosophy [φιλοσοφεῖν] by fleeing for refuge into argument [ἐπὶ δὲ τὸν λόγον καταφεύγοντες] and that in this way they will be good [σπουδαῖοι]. They act in a respect similar to those sick people who listen carefully to doctors, but then do none of the things that are prescribed. Thus, just as those people will not be well in body by treating themselves in this way, neither will the other group [be well] in soul by philosophizing in this way. (1105b12-8)

Now where did the many get this idea? We can find the target for Aristotle's criticism presented by Socrates himself in the heart of the *Apology*.

> If on the other hand I say that this happens to be the greatest good [μέγιστον ἀγαθόν] for human beings: to engage in discussions [λόγοι] every day about virtue and the other things about which you hear me talking and examining myself and others [. . .]. (38a1-5)

The greatest good for human beings is "to make *logoi*." *Logoi*, at their best, might result in conclusions and knowledge. This is deeply flawed by Aristotle's lights. To make people good they must do virtuous actions over and over, not have discussions about virtue. Compare the view expressed in Plato's *Charmides*. Charmides, the beauty who is suffering from a "mysterious" headache in the morning (and who later helps his uncle in a violent overthrow of the Athenian democracy), would like relief from his headache and Socrates would like to talk to him. So Socrates and Charmides' uncle, Critias, come up with a ruse that leads Charmides to them (155b-156a). Socrates will give him a "drug" (φάρμακον) for his headache, but in order to be effective, it must be administered together with "charms" (ἐπῳδαί). And what are the "charms" for our over-indulgent friend? A *logos* about temperance.

---

all described as "knowing." On McDowell's view, they cannot match precisely; the person with practical wisdom must "perceive" something different.

After explaining that one must treat not only the body, but also the soul, Socrates continues (supposedly relating the advice of a Thracian doctor, Zalmoxis):

> The soul is treated by certain charms [ἐπῳδαῖς τισιν], and the charms are these arguments [λόγοι], fine ones. And from such arguments temperance arises in souls, and when [temperance] has arisen and is present [in the soul] it is then easy to provide health both for the head and for the rest of the body. (157a3-b1)

Socrates presents a view plainly opposed to Aristotle's. He does, at least, address the very question Aristotle claims is "most valuable" in the *EE*, and which Aristotle accuses him of neglecting: how does virtue arise? But according to Socrates what makes a person temperate is "fine arguments."[26] The *logos* that Socrates proceeds to have with Charmides concerns, of course, the question "what is temperance?". In some way not elaborated here, Socrates appears to believe that engaging in an argument about what virtue is, helps to make one virtuous. The *logos* will be successful if the answer to the "what is F?" question is discovered. Aristotle rejects this. Socrates is wrong to spend all of his *logoi* on the question of what virtue is, since knowing what virtue is is not the goal of ethics; rather it is to be virtuous and act virtuously. Furthermore *logoi*, apparently of any type, do not make people good; practice and habituation do. Charmides, then, does not need arguments. He needs to acquire the habit of not drinking so much, which he will acquire not by argument, but by, repeatedly, not drinking so much. Socrates, on Aristotle's account, is doing Charmides a disservice.[27]

5. Aristotle's "What Is Virtue?" Question

Having already explained how virtue arises in a person – by habituation – in II.5-6 Aristotle supplies a definition of virtue of character by providing the genus and differentia of it. He answers a question of precisely the same form as Socrates': "what is virtue?" (1105b19).

---

[26] This emphasis on *logoi* might make sense of why MM3 above says that Socrates believes that virtue is *logos*.

[27] It is by no means clear that Plato disagrees: we might interpret the *Charmides* as a display of how Socratic *logoi* fail to instill the qualities they hope to. I cannot develop this point here.

> Virtue then is a state concerned with decision, being in a mean relative to us, determined by reason and in the way in which the person of practical wisdom would determine it. (1106b36-1107a2)

Insofar as he has successfully provided an account of what all virtues of character have in common, namely being states of this sort, he has made a step towards answering Socrates' "what is F?" question. In the accounts of the specific virtues that follow Aristotle makes sure to show, rather tediously, that each of them fits with this definition, even if, in some cases one extreme or the other does not have a name. In fact, disappointingly from some perspectives, this is *all* he seems to argue for in his account of the particular virtues. He shows that were old Socrates to submit his definition to an elenctic examination and attempt to come up with a counterexample either by showing that there is indeed some virtue of character that is not such a mean state (the "too narrow" objection), or that there is something that is such a state but not a virtue of character (the "too broad" objection), he would fail.[28]

But in another sense Aristotle's "definition" is notoriously frustrating, and we can easily imagine that it neither would nor ought to satisfy Cleitophon from Section 2. If we are looking for a way of determining, in some particular instance, what the virtuous action is, this definition will be of little help. This is especially so once we realize, as Aristotle is careful to point out, that the mean is "relative to us" (1106b*ff*). We ought to be angry not a certain antecedently specifiable amount, but "at the right times, in regard to the right objects, towards the right people, with the right aim, and in the right way" (1106b21-3). Of course, it would be absurd to attempt to determine some amount of anger that is right in all circumstances, but it seems that Aristotle has bought himself a plausible account at the cost of making it trivial. How much anger should I express in some particular circumstance? Well, the right amount. What is the right amount? As much anger as the person of practical wisdom would have shown. Socrates' desire expressed in the *Euthyphro* to have some way of looking at particular actions and determining whether the token action counted as an action of this type is certainly not satisfied by Aristotle's "definition." Insofar as this is the case, then, Aristotle may have supplied an answer of sorts to Socrates' "what is F?" question, but he has not provided a Socratic definition.

So far I have made a familiar criticism. Aristotle has not provided an agent with a way of determining what the virtuous course of action is in

[28] It is perhaps worth noting that even Aristotle's definition "works" only for the virtues of character, while Socrates was trying to come up with a definition for all virtue.

some circumstance. As Cleitophon complained, neither did Socrates. A Socratic definition *would* have provided such an account by providing, in effect, a universal principle against which token actions could be assessed and their virtue or lack determined. But Socrates notoriously fails at this, and now Aristotle seems not to be doing so well either. First Aristotle rejects Socrates' emphasis on acquiring an account of virtue that would give a person knowledge of what virtue is. Then, when he finally gives an account himself, it seems useless insofar as one wants to determine whether some particular action is the virtuous one or not.

So much for what Aristotle has *not* done. Has he accomplished anything useful here? He certainly seems to spend a fair amount of time on what has been dubbed, "The Doctrine of the Mean," which would seem particularly odd given how it seems to fail so utterly when it comes to solving the central problem that a Socratic definition was supposed to solve: how do I determine whether this token action is what ought to be done or not? What if, however, we take his criticism of Socrates' search for definitions more seriously, and do not expect him to be solving that problem, do not expect him to be offering what Cleitophon is seeking?

Let us think about aims. We learn in *NE* I.7 that eudaimonia, by definition the supreme and ultimate aim of a person's life, consists in rational activity according to excellence. Acting virtuously, on Aristotle' account, is what the activity of eudaimonia is. Socrates, as we have seen, thought that one ought always to aim at acting virtuously. Aristotle connects this conceptually with eudaimonia: eudaimonia is the activity of acting virtuously. Socrates stopped with the universal principle, SV, expressed in various ways, e.g., it is never right to do wrong. That principle is what I am taking to be the paradigmatic example of an aiming principle. Once Socrates has secured this principle, he reasonably proceeds to the next puzzle: now that I am committed to acting virtuously, surely I ought to determine what virtue is. It is with this further question that Socrates, according to Cleitophon, utterly fails. But Aristotle does not go down this dead-end road. He veers off in two different directions. First he replaces the importance of the question of what virtue is, with the question of how virtue arises. Second, his answer to what virtue is does not so much as attempt to supply a way of determining what the virtuous course of action is in any circumstance. Rather, it provides a more specific aiming principle. Aristotle shows us that to aim at virtue (at least for the virtues of character) is to aim at a mean state. This is a surprising fact. To be virtuous is to be the *best*: an extreme state. But, paradoxically enough, to be the best one does not aim at an extreme action. To achieve an extreme state one aims at a mean.

This seems unique to the virtues of character; it does not hold in other areas of life. To be the best runner, one runs the fastest; the strongest person is the one who can lift the most. But the best eater or drinker is not the one who is able to eat or drink the most, nor indeed the least.

Thought of as a further development of the aiming principles articulated in Book I, the Doctrine of the Mean becomes a plausible and reasonable doctrine, and, given its paradoxical nature – to become the best do not aim at an extreme, but at a mean – worthy of the attention Aristotle gives it. It fails so utterly to help to determine what the virtuous action is because, simply, that is not its point: it is an aiming principle. Determining what is virtuous is something that will be developed not by an account of virtue, but by habituation.

We have seen also that once Aristotle embarks on the topic of virtue in Book II, he does what he says one ought to do in the *EE*: he addresses the question of how virtue arises, rather than the question of what virtue is. When he comes to the critical question of describing what constitutes virtuous action, there is a focus on decision and on character, rather than on knowledge. The former arise not from teaching but from practice and habituation. Turning to the Socratic-sounding question of what virtue is, Aristotle answers in a way that specifies what all virtues (of character) have in common, but not in a way that would have enabled Socrates to do what he wanted to do with such an answer: determine which actions are virtuous and which are not. A Socratic definition, unlike an Aristotelian one, would have given us a general principle that could determine which actions were virtuous. Aristotle's "definition" is merely an aiming principle, telling us to aim at the mean relative to us. Determining the correct content of that principle is left to the properly brought up individual.

Furthermore, we ought to notice the striking absence of argument in the *NE* around the Socratic "what is F?" question regarding the particular virtues. Which are the virtues and how are they defined? Aristotle nowhere attempts to argue that his list of virtues is either correct or complete. Consider the way the particular virtues are introduced. In II.7 Aristotle provides an outline of all of the particular virtues, or as he has established by this point, "mean states," that he will consider. They include: courage, temperance, generosity, magnificence, magnanimity, virtue concerned with "small honors," mildness, friendliness, truthfulness, wit, shame (a sort of semi-virtue), proper indignation,[29] justice, and "the virtues of reason." I have listed all of the particular

[29] This is the only "virtue" not discussed again in the *NE*; but see *EE* III.7, and *Rhetoric* (II.6, 9-10) (I owe the references to Irwin [1999], note *ad loc*.).

virtues to emphasize the complete lack of any argument for the claim that these are the virtues, and the most important ones, and that they are to be described in this way. Aristotle does not begin his discussion of these particular virtues until III.6, introducing them at the end of III.5, as follows:

> Let us take up each [of the virtues] and say which they are and what sort of things they are concerned with and how they are concerned with them; at the same time it will become clear too how many they are. So first [let us speak] about courage. (1115a4-6)

After the discussion of courage ends in III.9, the next chapter begins: "after [courage] let us speak of temperance; for these seem to be the virtues of the irrational parts" (1117b23-4). This sentence, although containing a "for [. . .]" clause, argues only for the reasonableness of considering temperance after courage since both belong to the "non-rational parts" of the soul. It does not so much as raise the question of whether temperance and courage are in fact excellences; it simply appeals to the audience to concede that they have a similarity insofar as they are excellences of the non-rational part of soul. That Aristotle's account of the virtues is quite parochial and that he does not argue for their correctness is not a new idea in Aristotelian scholarship. But I would like to think about it in the context of the Socratic dialogues. When Aristotle provides no argument for his substantive account of the virtues, but merely "descriptions" of them, he in effect refuses even to attempt to answer Socrates' "what is F?" question. He has answered *a* "what is F?" question as we have seen in the previous section, but not Socrates'. A Socratic answer would have provided us with a principle for determining whether a token action is in fact, say, courageous (for futher discussion, see Vasiliou forthcoming, c. 4). There do not seem to be any such principles in the *NE*. There are, however, principles of another sort: aiming principles.

6. McDowell on Habituation, "Primitive" Phronesis, and Phronesis

In a series of papers, John McDowell has developed an important, if controversial, reading of Aristotle's ethics. The focus of McDowell's work is on Aristotle's moral psychology and, in particular, his conception

of *phronesis*.[30] I shall consider in this section how what we have seen Aristotle say about virtue and argument fits with and supports McDowell's understanding of *phronesis* and the role of habituation in its development. Then, in the final two sections, I shall argue that closer examination of the role that Aristotle says argument plays in moral development and of the actual place of argument in the *NE* itself indicates that there is more to moral development, and possibly to *phronesis*, than McDowell himself includes.

Consider the following passages from McDowell:

> Aristotle's picture is [. . .] this: a correct conception of how one should live does not yield a method of determining which concern one should act on; but one rather than another of the potentially practically relevant features of the situation would strike a virtuous person, and rightly so, as salient, as what matters about the situation. If there were 'ordering principles', they would yield an argument that what appeals to one rather than another of the concerns is what matters about the situation. In the absence of such an argument, it comes naturally to say 'You have to see it', with the perceptual content marking a point at which discursive justifications have run out (cf. *EN* 1143b1). (McDowell 1998, p. 112)

> Having the right end is not a mere aggregate of concerns; it requires the capacity to know which should be acted on when. If that capacity does not consist in acceptance of a set of rules, then there is really nothing for it to be except the capacity to get things right occasion by occasion: that is, the perceptual capacity that determines which feature of the situation should engage a standing concern. (McDowell 1998, p. 113)

McDowell is concerned with how the *phronimos* acts in the here and now. His answer is that the ability of the *phronimos* to determine what the right thing to do does "not consist in acceptance of a set of rules," and so "there is really nothing else for it to be except the capacity to get things right occasion by occasion." In the dialogues of definition, an answer to Socrates' "what is F?" question would have provided the agent with a rule that could have been used to answer the *phronimos*' question about the here and now. McDowell's account attempts to explain how the *phronimos* answers the question, "what is the virtuous action here and now?", given that he *lacks* any such general definition or rule. On McDowell's view it is upbringing and habituation that provide the *content* of one's moral outlook:

---

[30] I shall simply write '*phronesis*' for φρόνησις; translations include 'practical wisdom' (Oxford translation) or 'intelligence' (Irwin 1999). The "*phronimos*" is the person who possesses *phronesis* – the ideal Aristotelian moral agent.

> [. . .] what determines the content of a virtuous person's correct conception of the end is not an exercise of practical intellect, but rather the moulding of his motivational propensities in upbringing, which is described in book II of the *EN* as instilling virtue of character. (McDowell 1998, p. 114)

> I think Aristotle's view is that it is the moral development effected by upbringing that puts us in a position to undertake ethical deliberation. His account of the habituation that sets up states of character already contains enough to display states of character as having the intellectual aspect that he insists on. *If the content of a correct conception of doing well is fixed by proper upbringing*, that renders it superfluous to credit that role to an autonomous operation of the practical intellect, or to look to the intellect for a foundation for the claim that this rather than that conception of doing well is correct. (McDowell 1996a, p. 19, my emphasis)

We have seen that Aristotle's account of habituation explicitly replaces Socrates' "what is F?" question. Aristotle does not so much as attempt to answer Socrates' "what is F?" question, because he rejects the idea that one is going to find the content of virtuous action there. One must ask, rather, how virtue arises. This fits well with McDowell's contention that it is upbringing and habituation that provide the *content* of the virtuous person's correct conception of the end.

Particularly in (1996a), McDowell's account has habituation play an enormous role:

> The relevant habituation includes the imparting of conceptual apparatus, centrally the concept of the noble. That concept crystallizes the pleasure that an agent has learned to take in certain actions into the form of a reason for undertaking them. *The ability to see actions as noble is already a perhaps primitive form of the prescriptive intellectual excellence, practical wisdom, with its content intelligibly put in place by habituation.* (p. 28, my emphasis)

Phronesis, on this interpretation, is almost entirely effected by upbringing. McDowell does not say much about what upbringing itself involves, but it results in a double-sided ability to discern correctly what the right thing to do is in a particular situation and the motivational propensity to do it. Central to his account of Aristotle's moral psychology is that these are not two separate features of the *phronimos*, developed in different ways (see Whiting 2002a). Moreover, they are the upshot of one and the same process: habituation. McDowell maintains that the intellectual excellence, *phronesis* (or at least a "primitive" form of it), results from the formation of character "out of psychic materials that, before the formation of character, are not a source of rational

prescriptions" (McDowell 1996a, p. 27). So proper habituation yields a person who can both correctly identify what doing well is in the here and now, and who "delights" in doing it.

McDowell explains that although habituation provides the content of virtue, we should not fear that any upbringing is, rationally speaking, on a par with any other. If this were the case, there would be no way of establishing the objective correctness of Aristotle's substantive account of the virtues, which is something he clearly wants to do. McDowell diagnoses the source of this concern about how to establish the objectivity of an ethical position as part of a modern, and philosophically questionable notion of objectivity.[31] Freed from a mistaken and fortunately optional desire to validate one's ethical position "from outside," we can replace it with "Neurathian reflection," reflection that takes place piecemeal from within the inherited ethical framework. Although such "validation" can look like a lame second best, he believes that what it appears to be second to is no more than a philosophical fantasy. Neurathian reflection, then, turns out to be second to nothing; such reflection is the only support for a substantive moral outlook there is.[32]

McDowell stops short of saying that habituation by itself yields full-fledged *phronesis*. The intellectual state yielded by upbringing is merely "primitive" because all that such a person has is "piecemeal deliverances" of the state, which McDowell connects to having the "that" (cf. *NE* I.4, 1095b4-8). But such a person has not yet reflected on how these piecemeal applications hang together. If one engages in Neurathian-type reflection (which McDowell takes to involve the transition from possession of the "that" to the "because"), then "intelligibility accrues to the parts from their linkage into a whole." He says further that:

> [Aristotle] proceeds *as if the content of the conception of doing well is fixed once and for all*, in the minds of the sort of people he assumes his audience to be, by their upbringing; as if moral development for such a person is over and done with at the point when his parents send him out into the world to make his own life. There is no suggestion that an increase in reflectiveness and explicitness will alter the substance of the conception. (McDowell 1996a, p. 31, my emphasis)

---

[31] See McDowell (1998, pp. 116-119). This is a broader theme of McDowell's work, beyond ancient philosophy: see McDowell (1979) and (1996c). In Vasiliou (1996, §8), I discuss Aristotle's conception of moral objectivity in light of McDowell's.

[32] I am condensing a large amount of material and argument in these sentences. Also I am clearly only stating what is certainly a contentious view, not offering arguments for or against it. For further discussion, see Whiting (2002b).

McDowell is being somewhat critical of Aristotle here. He claims that nothing prevents a person from reflecting in an Aristotelian spirit so that the substance of the conception might be altered by modifying its content or perhaps even by creating new thick ethical concepts to deal with novel circumstances. Thus Aristotle's theory is not necessarily objectionably conservative and rigid; it leaves room for Neurathian reflection that can expand and develop one's moral concepts without thereby committing the alleged mistake of attempting to validate and justify one's ethical outlook from some standpoint that is independent of that outlook. McDowell doesn't say it in so many words, as far as I know, but it is in this reflection, if anywhere on his reading, that something recognizable as argument comes into play in the moral development of a person. On Aristotle's own account, according to McDowell, the "content of the conception of doing well is fixed once and for all [. . .] by upbringing" (McDowell 1996a, p. 19; see also quoted passage above). McDowell believes that Aristotle himself does not go much further than this, but tries, on Aristotle's behalf, to equip him with the final step of reflection as necessary before a person can be said to have full-fledged *phronesis*. In critical reflection, an uncritical example of which occurs in Aristotle's character sketches, the almost-*phronimos* would presumably engage in argument about how to incorporate new cases which do not fit as they are into his inherited moral outlook. He would reflect on similarities and differences between cases and in this way begin to expand his ethical concepts. This is something that Aristotle himself does not do much, but a good Aristotelian could, and presumably should.

7. Building *Logos* Up: Argument in the *NE*

We have seen that on McDowell's understanding of habituation, a proper upbringing supplies almost full goodness: a "primitive" *phronesis*. What is left to be done is to reflect on one's inherited moral outlook and concepts. Such reflection will provide the transition from proto-*phronesis* to the full-fledged state. Argument, other than Neurathian reflection on one's inherited ethical beliefs (something the *NE* itself does quite rarely), plays little or no role for McDowell's Aristotle. While McDowell's understanding of *phronesis* is quite controversial,[33] it appears to gain support from the dearth of argument about virtue that we discussed above

[33] See Irwin (2000) for recent criticism. Whiting (2002b) provides a thorough discussion of the differences between Irwin's and McDowell's positions and argues for a reconciliation.

(§5). When it comes time to list particular virtues, out comes "the diagram" (1107a32-33; διαγραφή); there is no argument offered about its correctness or completeness. Furthermore, one can agree with McDowell that there is no way to get from the relatively "thin" content of something like the Function Argument, which *is* an argument, to Aristotle's own substantive account of the virtues.[34] But there is another aspect of Aristotle's account that does not concern establishing the content of *virtue*, but the content of *eudaimonia*. Given Aristotle's account of eudaimonia these two issues inevitably overlap, but they are not identical. Even if there is no argument in the *NE* about what *virtue* is, there are many passages clearly recognizable as argument about what *eudaimonia* is – that is, which seek to establish the importance of having and aiming at an end, and of providing some specification of it. McDowell may be right that such an argument would have little content, and thus be of little use, without the upbringing that supplies the substantive content of virtue. But there remains a marked contrast between the *lack* of argument about virtue, and the dense *presence* of arguments around the question of what eudaimonia is. And these latter arguments are distinct in type from the arguments McDowell believes are part of the Aristotelian picture – namely, Neurathian reflection on one's acquired ethical beliefs.

From the opening of the *NE*, Aristotle provides arguments for the ideas that there is a highest good, that that good may be an activity (I.1-2), and that it must be most complete and self-sufficient (I.7, 1097a15-b21). He provides arguments for why eudaimonia ought not to be considered simply pleasure, honor, mere possession of virtue, luck, or money-making (I.5, I.8). In I.6, Aristotle provides a series of dense, shorthand arguments for why there can be no Form of the Good. He argues that the highest good can be further specified by consideration of the human function, which leads to the conclusion that eudaimonia centrally involves virtuous activity (I.7, 1097b22-1098a22). Finally, as we have seen, Aristotle argues that for the virtues of character, at least, "what virtue is" is a mean relative to us (II.6). The contrast between the quantity of what are clearly *arguments* on these topics versus the absence of argument when one wants to determine which virtues and which actions are actually the right ones is striking. I intend this (incomplete) list of arguments to highlight the fact of their presence and to call attention to what they are about, with a view to contrasting this with the absence of argument regarding determining what virtue is. As we have

[34] See McDowell (1980) and additional references and discussion in Vasiliou (1996, §4).

seen, in the latter cases Aristotle turns to *description* rather than argument.

This difference between the topics which receive argument in the *NE* and those which do not matches the distinction between the idea of a having virtue as an *aim* and the subsequent attempt to determine what virtue *is*. We saw that Cleitophon is persuaded by Socrates to have virtue as an *aim*, while he complains that he is left utterly ignorant with respect to determining its *content*. Cleitophon sees these two questions as distinct, and I think Aristotle does as well. By not recognizing the distinction between questions about aim and questions about determination of content we risk missing an important aspect of Aristotle's ethics. McDowell is right that without the proper upbringing that supplies the content, the arguments are worth little or nothing – indeed this is just what Aristotle says in the passages where he requires a person to be well brought up before attending the lectures and where he disparages the idea that argument alone can make people good (see §2). But we should be careful not to conclude from this that argument has an altogether insignificant role in moral development.

Arguments about the nature of *phronesis*, incontinence, friendship, or virtue are not found outside of the ethical works. But arguments about what I have called "aiming questions" are found throughout Aristotle's *Protrepticus*. Although I cannot discuss the reconstruction of the *Protrepticus* or its position in the Aristotelian corpus here, the argument thus far might yield some potentially interesting conclusions. The fact that there are arguments in the *Protrepticus* that closely match those in *NE* I and X.7-8 about the nature of eudaimonia and the ultimate goal of a human life supports the idea that the arguments in these parts of the *NE* are distinct in some way; they may perhaps be fairly called "protreptic arguments."[35] But matters are complicated by the way in which these arguments have been contextualized in the *NE*. Although it is disputed whether protreptic represented a distinct genre in the fourth century (the evidence may well be too scant to determine this decisively) (see Slings 1999, pp. 59-93), in an example of it from Plato's *Euthydemus* Socrates is depicted as *beginning* his association with the boy Clinias via protreptic. From one perspective, this seems entirely reasonable: the first thing a person needs to do is to convince someone that she ought to pursue philosophy (and virtue). Then, once a person has been convinced via *logos*, she will begin with the actual activity of learning what virtue

[35] The fragment references to the Protrepticus refer to Düring's edition (1961): B11-B16/ I.9; B28/I.7 (1097b33-1098a5); B42/I.2 (1094a18-24); B52/I.8 (1095b31-1099a7) and II.2 (1103b26-30); B63-B68/X.7; B78-B90/I.7 (1097b22-1098a18).

and philosophy is. On this understanding of protreptic, it comes *first*, and then worry about the content involved in either philosophy or virtue comes afterwards. This is also in line with Cleitophon's complaint: the protreptic went fine – he was thoroughly convinced that he ought to pursue virtue – but there was no second course. I believe that Aristotle turns this around: he takes "protreptic arguments" seriously, but he limits their effectiveness to those who have *already* engaged in proper practice; that is, those who have been properly habituated. Thus for Aristotle, at least at the point of his writing the *NE*, protreptic is not something that occurs prior to learning the content of virtue.[36] Contrary to the impression that an isolated reading of the *Protrepticus* might give, however, these arguments are not simply valuable without qualification for anyone, anytime.

We should also recognize that what arguments in the *NE* and *Protrepticus* do not do is to supply a certain type of "because": why is giving this person ten dollars the virtuous thing to do here and now? This is not the sort of thing that is the subject matter of arguments in the *Ethics* or *Protrepticus*. Rather, they concern the establishment of aiming principles and associated concepts: that there is a highest good, that it ought to be self-sufficient, centrally involve virtuous activity, and so on. None of this will be helpful towards solving substantive deliberative problems concerning determining what doing well *is* here and now, nor is it meant to. But such arguments are intended to show someone who already possesses the *content* of virtue via his upbringing why the sort of life that he has been specially brought up to value is worth standing by and clinging to. These arguments treat aiming questions and offer justifications for why the well brought up person ought to embrace virtue and wisdom rather than, for example, pleasure, or financial gain.

But before examining the value Aristotle attributes to such arguments for moral development, let us consider the somewhat awkward position that he is in given his beliefs, discussed above: that actually becoming good is the goal of ethics and that the way one does this is through habituation. Aristotle is writing (and/or giving) lectures on ethics, and ethics is about becoming good; but listening to lectures and hearing

[36] I shall not hypothesize about whether the *Protrepticus* should be considered an "early" work, as some scholars have argued (e.g., Jaeger 1948). But if it were, one might posit an evolution in Aristotle's position: beginning with a more Socratic conception of protreptic – that one should approach the young first with protreptic arguments to convince them of the value of virtue and philosophy – his later view, represented and emphasized in the *NE*, would then be that such arguments still have a critical value, but only for a certain restricted class of people who have already been habituated in proper ways. I am sceptical, however, that such developmental hypotheses could ever be more than speculative.

arguments is not how a person becomes good. If ethics is about actually being good, and one does not become good by argument or lecture, then in some sense one cannot give lectures or offer arguments that are truly part of *ethics*. If one wanted to do actual ethics, he would engage in the habituation either of himself or of someone else. So, to describe what the *NE itself* is doing more accurately, we should say that it is talking *about*, not actually engaging in, ethics since the latter consists in engaging in a process that does not involve either lecturing or argument, but practice. Does the *NE* itself, then, have any role to play in actually becoming good, or is it simply a description of the sorts of activities that one would need to engage in if one were to become good?

This question is especially acute for a position on Aristotle's ethics like McDowell's, since it maintains that habituation does not simply work on one part of the soul, the non-rational, but by itself yields at one and the same time a shaping of both a person's desiderative and intellectual capacities.[37] The person who has undergone proper habituation emerges, as we have seen, with a primitive form of practical wisdom. If McDowell is right, then a question about the role of the *NE* becomes even more pressing. For if proper upbringing supplies a person with the content of virtue, the only thing left is critical reflection on her inherited moral outlook to complete the transition to full-fledged *phronesis*, and, as McDowell admits, the *NE* does not engage in much reflection of this sort. The arguments of the *NE* will still, of course, appeal to philosophers who concern themselves with getting the right account or theory of matters such as happiness, virtue, friendship, incontinence, pleasure, the value of external goods, and *phronesis*. But I shall note an obvious point: getting the right account of *what phronesis is* is not the same as *having phronesis*. The debate between, for example, McDowell and Irwin about the correct understanding of Aristotelian *phronesis* is a debate between two philosophers about a philosophical account. Aristotle too seems to think of himself as doing philosophy in the *NE* (1096b31, 1181b15). But does the *NE* have anything of value for the person who wants to become practically wise, and not just to know what practical wisdom is? On McDowell's reading the answer would seem to be "no."[38]

---

[37] Whiting (2002a) argues in detail that the *De Anima*, not considered by McDowell, may support his view that the desiderative and cognitive parts of the soul that are relevant to action (as opposed to nutrition or pure theoretical contemplation) are in fact unified.

[38] A second audience for the *NE*, besides philosophers, is certainly also the statesman. Broadie (1991) discusses this, see especially chapter one.

The so-called "good-upbringing" passages make it clear, however, that Aristotle is addressing not only philosophers or future statesmen. Uniquely in the corpus, Aristotle requires that the listeners be well brought up, which is explained as having acquired the right habits *before* listening to the lectures. Scholars have recognized that this is because Aristotle does not believe that the corrupt will understand or get benefit from these lectures. As we have discussed, bad people cannot be made good by argument (see §2). McDowell can take this aspect of these passages in stride, since he thinks it is habituation that inculcates almost full practical wisdom. But they do not simply support his position, for they show that Aristotle also believes that argument has a significant role to play in a person's *becoming* good. We are not then simply arguing amongst ourselves and with Aristotle about the philosophical question of *how* a person becomes *phronimos* and what the correct account of *phronesis* is (although we and Aristotle are certainly doing that). Aristotle is also saying that argument (philosophy) has a significant role to play in providing us with some sort of knowledge that will be a part of our moral development. What sort of knowledge is that? And is there any room for such a thing on a McDowellian account of Aristotle's ethics?[39]

[39] A further question arises about the genesis of the *NE* itself on McDowell's reading: when Aristotle composed it, which Aristotelian faculty was at work? McDowell (1998, p. 115), describes the position he *opposes* as follows: "There is a philosophical motivation for a reading of Aristotle in which what determines the content of a virtuous person's conception of the end is an exercise of the intellect." We might continue his sentence as follows: "[. . .] and insofar as the *NE* presents some of the content of the virtuous person's conception of the end, the *NE* itself must be the result of an exercise of practical intellect." On the view that McDowell is setting himself against, then, the *NE* is in part a philosophical account of what it is to have *phronesis but also* partly an exercise of that very faculty insofar as it presents the content of the good life as determined by that faculty. But if, on the contrary, McDowell is correct, the *NE* itself has nothing to do with the faculty of *phronesis* (that is, Aristotle is not displaying any *phronesis* in writing it). For any general content the *NE* possesses is not the result of a genuine exercise of *phronesis*, but simply summaries of particular actions of *phronimoi*, with which Aristotle can count on his well brought up audience to agree. I think this brings out an aspect of McDowell's difference with other commentators that is not typically highlighted. Not only does he present a differing account of *phronesis*, but his interpretation also has the additional consequence that *phronesis* has no role to play in the writing of the *NE*. On an account of *phronesis* such as Cooper's (1975/1986) or Irwin's (1988a) – despite great differences between them – the *NE* will itself be displaying part of the *content* of *phronesis* which is itself *effected by phronesis*, and so ought to be understood as an at least partial writing down of an exercise of *phronesis*. I cannot pursue this difference further here. See Irwin (1999, pp. 326-327), under 'ethics', and Irwin (2000) for criticism of McDowell's position.

In the final section we shall see that Aristotle dictates a role for argument in moral development above and beyond habituation.

8. The Importance of Argument for Moral Development

Not only does argument have a significant presence in the *NE*, Aristotle also claims that it has a significant role in moral development. While Aristotle criticizes the idea that argument is what *makes* people good (as we saw above), he also, in apparent tension with this, lauds argument and the knowledge it can yield. I shall claim that although the arguments of the *NE* do not help to determine the content of virtue, they are important for properly brought up agents deliberating about *aiming* questions. At the very beginning of the *NE*, having raised the possibility of there being a highest good in life, he says:

> Then surely knowledge [γνῶσις] of it [the best good] has a great weight [μεγάλην ἔχει ῥοπήν][40] in life [πρὸς τὸν βίον], and, just like archers who have a target [σκοπὸν], we would more likely hit upon what is right [τοῦ δέοντος]. (I.2, 1094a22-24)

What is this knowledge of the good? It cannot be the knowledge of the *phronimos*, for *phronesis* does not simply have a "great weight" or a large role to play in life: it is what is both necessary and sufficient for being good. Aristotle refers here to ordinary knowledge – γνῶσις. Further, Aristotle specifically focuses on the concept of an aim or target, as the archer image emphasizes. Knowledge of the aim can have a great role to play in life, or "towards living." It is not that such knowledge is by any means sufficient for being good. Barely a page later Aristotle will restrict his audience to those who are well brought up, saying that the goal of lectures on politics and ethics is not knowledge, but action, and dismissing the young as inexperienced in the actions of life and as tending to follow their passions instead of the results of argument (1094b28-1095a6), which is a clear reference to the necessity of proper habituation, and connects to passages we shall consider below from X.9. But the above passage also highlights the importance of knowledge (even though the *end* is action), and by implication, the importance of the arguments that will supply that knowledge, for a person who has undergone the appropriate upbringing. Aristotle closes his remarks in I.3 about the student of ethics by saying:

---

[40] 'ῥοπή' can mean "the weight which makes the scale turn," i.e., the *decisive* weight.

> For to such persons [those who are young in age or youthful in character], knowledge [γνῶσις] is useless [ἀνόνητος], just like to those who are incontinent; but to those who make their desires according to reason [κατὰ λόγον] and act [this way], knowing about these things [τὸ περὶ τούτων εἰδέναι] would be of great benefit. (1095a8-11)

Here Aristotle is still marking out a role for argument for one who has been properly habituated, as opposed to those who have yet to receive (or who have never received) proper habituation. If McDowell is right, such a person has a primitive *phronesis*. But the role for argument that Aristotle is pointing at concerns the *aims* of life: that is, the sorts of argument in which the *NE* does explicitly engage – and not the Neurathian reflection, in which the *NE* does not engage, that McDowell envisions as providing the further step to full-fledged *phronesis*. What is this other role that Aristotle sees for argument?

In the first chapter of Book X, where Aristotle revisits the topic of pleasure, he indicates a role for argument, and the knowledge that can be derived from it. He establishes the importance of discussing pleasure – because pleasure and pain touch almost all aspects of our lives – and then says that some people believe that pleasure is the good, while others believe it is altogether base. The latter group then subdivides into two: first, those who really believe that pleasure is base; and, second, those who believe it is better to *present* pleasure as wholly base, even though it is not. The second group justifies their purposeful use of falsehood on the grounds that most people are slaves to pleasure, and so by pushing them in the entirely opposite direction, they are most likely to hit the right mark. Aristotle believes that this is a bad strategy, however, for people will see that the actions of those who speak in this way conflict with their deeds, and this will create contempt both for arguments and for the truth. He then discusses the value of true arguments:

> Therefore true arguments [οἱ ἀληθεῖς τῶν λόγων] seem not only to be most useful [χρησιμώτατοι] in regard to knowing [πρὸς τὸ εἰδέναι], but also in life [πρὸς τὸν βίον]. For since [true arguments] harmonize with deeds [τοῖς ἔργοις] they carry conviction, and for that reason [διό] they encourage [προτρέπονται] those who comprehend [τοὺς συνιέτας] [them] to live according to them. (X.1, 1172b3-7)

In the earlier passage (1094a22-4) we saw that Aristotle also said that knowledge of the highest good is beneficial "in life [πρὸς τὸν βίον]." The earlier implicit connection is here made explicit. True arguments, which presumably yield knowledge, are also useful in life. They are important because they can "encourage those who understand them to live according to them." Aristotle uses the verb προτρέπω. Cleitophon lauds

Socrates for his protreptic skills. But Aristotle, while praising the importance of protreptic argument, restricts its usefulness to "those who comprehend them." *For such types* – the well-brought up who have received proper habituation and have the right habits – arguments are of practical significance. As additional passages in X.9 confirm,[41] Aristotle maintains that *after* a certain amount of habituation, argument has an important practical role to play.[42] What is this role, and does it fit with a McDowell-type understanding of *phronesis*?

In the final chapter of the *NE* (X.9), Aristotle elaborates on the role and limits of argument for ethics and politics.

> If arguments [λόγοι] were sufficient [αὐτάρκεις] for making people decent [ἐπιεικεῖς], they would, according to Theognis, justly bring numerous and great rewards, and these [rewards] ought to be furnished. In fact [νῦν δὲ], however, they appear on the one hand to have the power to encourage [προτρέψασθαι] and to stimulate [παρορμῆσαι] the deservedly free of the youth [τῶν νέων τοὺς ἐλευθερίους], and to make a person of noble character and a true lover of the fine capable of being possessed by virtue [κατοκώχιμον ἐκ τῆς ἀρετῆς], but on the other hand [arguments] are unable to encourage [προτρέψασθαι] the many towards being noble-and-good [πρὸς καλοκαγαθιάν]. For [. . .] [what follows is a long explanation of why arguments alone do not work, and do not work on the many]. (X.9, 1179b4-10)

What sort of arguments is Aristotle talking about? I would claim that they are not arguments about what to do in the here and now, for example, whether or not to have a fourth drink, or whether or not to help a friend. They are, rather, arguments about what sort of life to lead. Why do I say this? The passage begins by challenging the idea that arguments are sufficient to "make people decent." To make a person decent is to determine what kind of person he is. Aristotle has already told us that habituation is what does this, and he repeats this throughout X.9.

---

[41] See, for example, 1179b23-26, 29-31: "Argument and teaching do not prevail in all cases, but it is necessary that the soul of the listener be prepared beforehand [προδιειργάσθαι] by habits for delighting in and hating [things] finely [. . .] it is necessary that the character already be somehow [προυπάρχειν πως] fitted to [οἰκεῖον; "at home with," as we say in English] virtue, taking joy in the fine and hating the shameful." Note the clear temporal order, in part conveyed by the prefixes προ-. In order for protreptic argument to be effective, the student must be habituated first.

[42] Whiting (2002b, pp. 95-98), reads Aristotle as claiming that protreptic arguments ought to come first, and so as not necessary for people who have already received proper upbringing. On her account protreptic arguments are useful to motivate people who need motivating to seek out "the sort of habituation that is required to *complete* the process of becoming genuinely virtuous [. .]" (p. 98, bottom, her emphasis).

Protreptic argument is not about encouraging any random person to do or not do *some particular action*, nor does it teach anyone what virtue is in general. Rather, in the philosophical context, protreptic argument encourages a certain type of person to lead a certain kind of life. For the proper subjects – ones who are noble in character and true lovers of the fine, that is, the well brought up – argument has the power to make such a person "capable of being possessed by virtue." This is precise language. What does Aristotle mean by saying that protreptic argument makes the right subject able to "be possessed by virtue?" Once a person has been properly habituated, and acquired the right habits, they have, as McDowell believes, acquired the ability to discern what the right thing to do is in particular circumstances and they desire to do it. In short, they have been habituated into the *content* of virtue. Once the young man has this upbringing, however, so that he will, as part of his "second nature," do and take pleasure in what he ought, he is still in need of something further: argument. Thus, the *NE* has a role to play not just theoretically for philosophers, but practically for a person who has been properly brought up and wants to become fully good.

Such a person not only needs to know which actions are virtuous and be motivated to perform these (which he has from his upbringing) but he then also needs argument that can make him be possessed by it – that is, make him committed explicitly to living a life of virtuous activity, as opposed to a life of pleasure or money-making or honor. Of course, he will already be disposed as a result of his upbringing to value virtuous action and to act virtuously, but we can see this as a result of, in part, his appropriately restricted experience. Given his good upbringing, he has never been challenged by people proposing alternative conceptions of the good. It is true that he will not, initially, be disposed to act in ways contrary to virtue. But as X.9 makes clear, habituation is not a process that comes to an end. We saw that McDowell believes that Aristotelian habituation "fixes the content of doing well once and for all." But in X.9, when Aristotle is preparing a transition from ethics to politics, he argues for the importance of a state's having the proper laws – something he started to discuss back in II.1 in the initial account of habituation (1103b2-6). There is an important connection between proper laws and habituation. Aristotle emphasizes it because a person is continually affecting his character by the actions he engages in.

> It is not enough indeed that that those who are young happen to get correct nourishment and care, but it is necessary as well to practice these things and to be habituated once people are men, and we need laws for these things, and generally for all of life. (1180a1-3)

Childhood and adolescent habituation does not fix things once and for all; it is a relatively early, if extremely important, stage in moral development. Habituation continues to affect a person's character at any point in his life – although, as we have discussed, habits are not easy to change once acquired. Just as a bad person may become good, by changing the ways he acts and talks, there is no reason to deny the possible demise of a well-brought up person by coming to act in the wrong ways. In particular, we can think of an adolescent, well-brought up by his parents, who is suddenly exposed in the world to new and vicious ways of acting. One's first cigarette, first sip of scotch, and so on, are rarely initially pleasant; one begins on a new road of habituation. Faced with temptations (which, admittedly at first will not be tempting, given his good upbringing) and arguments that advocate, say, excess (for example, a drinking or eating contest), a person who wants to remain excellent, who wants not merely to know and be motivated by virtue, but "to be possessed by it," will need arguments to understand why the life he has been brought up to lead is indeed the best. Arguments can help to ensure that the *content* of virtue holds the place in a person's life that it ought to. One is continually undergoing habituation insofar as one is continuing to engage in actions of a certain sort. That is why moral change is possible for many, perhaps for almost all. That is why, too, politics and appropriate legislation are necessary. When parents send their child into the world her initial habituation may be complete, but she is not "safe": she will need *logos*, protreptic *logos*, to make her "capable of being possessed by virtue." Only this *logos* will keep the content she has learned through habituation aimed in the right way.

The arguments I have called attention to, and my interpretation of how to understand them, do not attempt to establish the correctness of Aristotle's substantive conception of virtue from a position outside of the inherited outlook. It has been Aristotle's repeated concern to emphasize that the arguments alone will be empty and useless. To know that eudaimonia is rational activity in accordance with virtue, or that virtue consists in a mean relative to us, is impotent without virtue having some *content* – a version of Cleitophon's complaint. That content is to be supplied, as we have seen, by habituation. In agreement with McDowell, there is no argument whose conclusion *is* that content. But there are other arguments which show why a *commitment to* that content is right; there are arguments that can help a person with that content to be possessed by it, to hold on to it in the face of competing life aims. Such arguments do not provide any sort of *foundation* or *justification* for the correctness of the substantive moral outlook. The content is always simply *described*.

Rather, the protreptic argument says: given that content, which I can loosely summarize, here is a reason why one should aim at that sort of life, rather than other sorts. Aristotle presents such arguments, especially in Books I, II.6, and X.7-8, and thinks that they have a critical role to play in ethics. I have tried to show, therefore, that Aristotle does believe that the *NE* and its arguments have practical relevance for being good by providing protreptic to those who are well brought up.

We now recognize that Socrates' problem, by Aristotle's lights, is that he engages in protreptic argument with those who are not in a position to be "stimulated and encouraged" by argument (for example, as is the case with Callicles) because they are not properly brought up. Without the *content* of virtue, which comes from habituation and *not* argument, argument is at best useless, and one ends up with the complaint of Cleitophon. But even though it is not argument that establishes the content of virtue, argument still has a role to play in keeping a person aimed in the right way, not in the face of a particular temptation or particular moral dilemma, but in the face of alternative ways of living one's life. Protreptic argument is about persuading a person to stay with virtue and philosophy, the content of which is established initially in upbringing. Since the relationship between actions and formation of character, described in II.1 and 4, holds forever, one's character is continually undergoing either reinforcement or change. Protreptic argument then provides a "because," but it is not the "because" generated by reflection on the particular deliverances of the perceptual capacity developed in upbringing, but by reflection on the aim of one's life as a whole.

Acknowledgements

I would like to thank John Partridge and Sergio Tenenbaum for comments on an earlier draft of this paper.

The Graduate Center/Brooklyn College
City University of New York
Department of Philosophy
2900 Bedford Ave.
Brooklyn, NY 11210-2889, USA
e-mail: vasiliou@brooklyn.cuny.edu

REFERENCES

Bekker, I., ed. (1831-1870). *Aristotelis Opera*. Berlin: Reimer. Translation in: Barnes, J., ed. (1984). *Complete Works of Aristotle*. 2 vols. Princeton: Princeton University Press.

Broadie, S. (1991). *Ethics With Aristotle.* Oxford: Oxford University Press.

Burnet, J., ed. (1902-1906). *Platonis Opera*. 5 vols. Oxford: Clerendon Press. Translation in: Cooper, J., ed. (1997). *Plato Complete Works*. Indianapolis: Hackett.

Burnyeat, M. (1980). Aristotle on Learning to Be Good. In: Rorty (1980), pp. 69-92.

Cooper, J. (1975/1986). *Reason and Human Good in Aristotle.* Cambridge, MA: Harvard University Press/Indianapolis: Hackett Press.

Di Muzio, G. (2000). Aristotle on Improving One's Character. *Phronesis* **45**, 205-219.

Düring, I. (1961). *Aristotle's Protrepticus*. Göteborg: Almqvist & Wiksell.

Engstrom, S. and J. Whiting, eds. (1996). *Aristotle, Kant, and the Stoics*. Cambridge: Cambridge University Press.

Heinamen, R. (1995). *Aristotle and Moral Realism.* London: University College London Press.

Irwin, T. (1988a). *Aristotle's First Principles*. Oxford: Oxford University Press.

Irwin, T. (1988b). Some Rational Aspects of Incontinence. *The Southern Journal of Philosophy* **27** (supplement), 49-88.

Irwin, T. (1995). *Plato's Ethics*. Oxford: Oxford University Press.

Irwin, T. (1999). *Aristotle Nicomachean Ethics*. 2nd edition. Indianapolis: Hackett Press.

Irwin, T. (2000). Ethics as an Inexact Science: Aristotle's Ambitions for Moral Theory. In: B. Hooker and M. Little (eds.), *Moral Particularism*, pp. 100-129. Oxford: Oxford University Press.

Jaeger, W. (1948). *Aristotle: Fundamentals of the History of his Development*. 2nd ed. Oxford: Oxford University Press.

Kahn, C. (1996). *Plato and the Socratic Dialogue*. Cambridge: Cambridge University Press.

McDowell, J. (1979). Virtue and Reason. *The Monist* **62**, 331-350.

McDowell, J. (1980). The Role of Eudaimonia in Aristotle's Ethics. In: Rorty (1980), pp. 359-376.

McDowell, J. (1996a). Deliberation and Moral Development in Aristotle's Ethics. In: Engstrom *et al.* (1996), pp. 19-35.

McDowell, J. (1996b). Incontinence and Practical Wisdom in Aristotle. In: S. Lovibond and S.G. Williams (eds.), *Essays for David Wiggins: Identity, Truth, and Value*, pp. 95-112. Oxford: Blackwell Publishers Ltd.

McDowell, J. (1996c). *Mind and World* (with a new Introduction). Cambridge, MA: Harvard University Press.

McDowell, J. (1998). Some Issues in Aristotle's Moral Psychology. In: S. Everson (ed.), *Companions to Ancient Thought 4: Ethics*, pp. 107-128. Cambridge: Cambridge University Press.

Nussbaum, M. (1990). The Discernment of Perception: An Aristotelian Conception of Private and Public Rationality. In: *Love's Knowledge*, pp. 54-105. Oxford: Oxford University Press.

Rorty, A.O., ed. (1980). *Essays on Aristotle's Ethics*. Berkeley and Los Angeles: University of California Press.

Sherman, N. (1989). *The Fabric of Character*. Oxford: Oxford University Press.

Sherman, N. (1997). *Making a Necessity of Virtue: Aristotle and Kant on Virtue*. Cambridge: Cambridge University Press.

Silverman, A. (2002). *The Dialectic of Essence: A Study of Plato's Metaphysics*. Princeton: Princeton University Press.

Slings, S.R. (1999). *Plato's Clitophon*. Cambridge: Cambridge University Press.

Vasiliou, I. (1996). The Role of Good Upbringing in Aristotle's Ethics. *Philosophy and Phenomenological Research* **56** (4), 771-797.

Vasiliou, I. (forthcoming). *Aiming at Virtue in Plato*. Cambridge: Cambridge University Press.

Whiting, J. (2002a). Locomotive Soul: The Parts of Soul in Aristotle's Scientific Works. *Oxford Studies in Ancient Philosophy* **22**, 141-200.

Whiting, J. (2002b). Strong Dialectic, Neurathian Reflection, and the Ascent of Desire: Irwin and McDowell on Aristotle's Methods of Ethics. In: J. Cleary and G. Gurtler (eds.), *Proceedings of the Boston Area Colloquium in Ancient Philosophy*, vol. 17, pp. 61-122. Boston: Brill.

Whiting, J. (2002c). Eudaimonia, External Results, and Virtuous Actions. *Philosophy and Phenomenological Research* **65** (2), 270-290.

Williams, B. (1995). Acting as the Virtuous Person Acts. In: Heinamen (1995), pp. 13-23.

Woods, M. (1982). *Aristotle's Eudemian Ethics Books I, II, and VIII*. Oxford: Oxford University Press.

Donald Ainslie

CHARACTER TRAITS AND THE HUMEAN APPROACH TO ETHICS

The revival of virtue theory over the past 25 years has meant that G.E.M. Anscombe's famous complaint that we are "conspicuously lacking" an adequate "account of what *type of characteristic* a virtue is [. . .] and how it relates to the actions in which it is instanced" (Anscombe 1958, p. 3) no longer seems warranted. But there remains at least one fundamental issue relating to the virtues that continues to be neglected, namely, how we are able to *recognize* one another as the bearers of character traits, at least some of which will qualify as virtues and vices.[1] Since this ability is central to our moral experience in everyday life – after all, we routinely use trait-terms to evaluate one another as, say, cheerful, generous, industrious, witty, fickle, or placid – it is worth pausing briefly to consider why its study has been neglected. The problem, I suspect, is that the approaches to moral theory that have been dominant of late, whether Aristotelian, Kantian, Hobbesian, or utilitarian, have all shared a focus on issues relating to practical reason. The primary concern has been the question 'What ought I to do?' and, despite the innumerable differences between the answers offered to it, the question 'How ought I to evaluate someone?' – the question of how we ascribe virtues or vices to a person – has been downplayed by all.[2] Kant, for example, goes so far as to say that one should avoid trait ascription altogether: Since the ultimate ground of human action is not part of the phenomenal world, we should respond to the appearances of behavior

[1] A few exceptions to this general neglect have been: Brandt (1970), Alston (1975), Sabini *et al.* (1982), and Butler (1988). As Alston, and Sabini and Silver point out, although traits have been a topic of some interest to personality theorists in psychology, their different concerns have meant that their treatment of the issue has shed little light on it from the point of view of philosophy.

[2] Robert M. Adams is a notable exception to this trend; see his (1985).

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 79-110. Amsterdam/New York, NY: Rodopi, 2007.

that we do have access to "with philanthropy [. . .], not merely by softening our judgements but also by silencing them" Kant (1983, p. 132 [Akademie edition, p. 466]).[3] Even the Aristotelians, who have devoted the most energy to a consideration of the virtues, have been more concerned to analyze the dispositions that amount to *phronesis* (practical wisdom) than to take up the issue of how we identify someone as virtuous or vicious in some respect.[4]

There is one tradition of moral thought, however, which puts the spectator's question – 'How ought I to evaluate someone?' – at the center, namely, the sentimentalism of the 18th-century British philosophers such as David Hume. And thus in my attempt to elucidate the practice of character-trait ascription, I will analyze Hume's treatment of this issue, primarily as it appears in his *Treatise of Human Nature*.[5] Hume repeatedly says that moral evaluation is of character traits, not actions (his argument for this claim will be the topic of §1), but despite this emphasis, his treatment of trait ascription is somewhat elusive. As Jane McIntyre has pointed out in one of the few articles to address this lacuna in Hume's philosophy (1990),[6] it might seem as though two of Hume's fundamental metaphysical and epistemological commitments leave him *unable* to offer a successful account of character trait ascription. First, his reduction of the self to a "bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement" (T 1.4.6.4, SBN 252) seems to leave him with no conceptual space for character traits, given that such traits persist in a person for significant portions of her life and are properly ascribable to her even when they do not issue in action (a point that Hume himself acknowledges when he says that "virtue in rags is still virtue" [T 3.3.1.19, SBN 584]). Second, his treatment of causation

[3] For a discussion of this feature of Kant's account of moral evaluation thought, see Korsgaard (1996).

[4] In the *Nicomachean Ethics* Aristotle does provide a brief treatment of the intellectual virtue of *sunesis* (translated by T. Irwin as 'comprehension'), which plays the role in our assessment of one another's virtue that *phronesis* plays in our practical reasoning (1985, Book VI, Ch. 10, 1143a7-15).

[5] Hume (1978), edited by L.A. Selby-Bigge and P.H. Nidditch; or (2000a), edited by David F. Norton and Mary J. Norton. Hereafter I will refer to the *Treatise* parenthetically as "T" followed by Book, Part, Section and paragraph numbers (as given in the Norton and Norton edition), followed by "SBN" and the page number as given in the Selby-Bigge and Nidditch edition.

[6] Further references to this article will be made parenthetically, using "McI" and the page number. In general the issue of character traits in Hume's theory has been remarkably underexamined, the only discussions of it occurring in: Bricke (1974), Baier (1991, Ch. 8), and Dees (1997).

as the constant conjunction of two types of events, observation of which conjunction leads onlookers to associate their ideas of these event-types (T 1.3.14.31, SBN 170), makes it hard to see how we can *ever* ascribe a character trait to someone. For, as Hume himself admits, we *never* observe such traits (T 3.2.1.2, SBN 477);[7] how then can we infer them from those parts of a person's behavior that we do observe?

McIntyre attempts to provide Hume with a solution to these apparent problems. She suggests that he sees a person's traits as the "relatively stable passions that give rise to a person's actions" (McI 201) and that this allows him to hold a realist account of traits in the face of his sceptical treatments of the self and of causation; we recognize such traits, on this interpretation, by a combination of causal reasoning and sympathetic feeling. After describing McIntyre's suggestion in more detail in §2, I will raise four objections to it in §3. My main concern is that McIntyre leaves Hume with no way of accounting for the *normativity* of character traits, the way in which they *demand of us* certain evaluative responses. In §4, I will offer an alternative interpretation of Hume's account of character traits; my suggestion is that the so-called *indirect passions* of pride, humility, love, and hatred play a crucial role in his theory of traits and trait recognition. In §5, I show how, armed with this conception of character traits, Hume is able to accommodate a certain kind of objectivity and normativity in his otherwise subjectivist and descriptive moral theory. I conclude in §6 by pointing to the consequences that follow from the Humean approach to character traits that I have excavated.

But, first, I shall explain why Hume thinks that the answer to the question 'How ought I to evaluate someone?' should involve an appeal to character traits.

1. Why Character Traits?

Hume first claims that moral evaluation focuses on a person's traits in his argument for a necessitarian position in the free-will debate. He points out that in order to hold someone responsible for some action that she or

[7] Even in our own cases, it is not clear that we can observe our character traits; and, if we could, it is not clear that our observation of the conjunction of these traits and our own behavior would be applicable to other people, given that the first-personal awareness we have of ourselves is so unlike the third-personal observations we make of others. Furthermore, given that we can recognize character traits in others that we ourselves do not have, Hume cannot rely on self-observation to explain our ability to ascribe traits.

he has performed, we must be able to connect the action to the person in question; it must be "infixed upon" (T 2.3.2.6, SBN 411) him or her. For an action, as a relatively short-lived event, is "by its very nature temporary and perishing" (T 2.3.2.6, SBN 411). If there is no sense in which the *person* who did the deed is somehow implicated in it, she or he cannot at a later time reasonably be held responsible for it. Hume suggests that a necessary condition for this implication or infixing, is that the action be derived from something "durable and constant" in the person, such as her or his "characters and disposition," her or his character traits (T 2.3.2.6, SBN 411; see also T 2.2.3.4, 3.2.1.2, SBN 349, 477).[8] Since these traits persist even while the action has been completed, the traits help to connect the action to the person.

This suggestion harks back to Hume's claim, a few pages earlier, that human behavior is the result of the interplay of the various kinds of durable dispositions persons possess:

> There is a general course of nature in human actions, as well as in the operations of the sun and the climate. There are also characters peculiar to different nations and particular persons, as well as common to mankind. The knowledge of these characters is founded on the observation of an uniformity in the actions, that flow from them; and this uniformity forms the very essence of necessity. (T 2.3.1.10, SBN 402-403)

Hume points here to three different causal sources for human actions: human, national, and individual "characters." In so far as we recognize the presence of such characters in another person (or even ourselves), we can infer the actions he is (or we are) likely to perform. The human character encompasses all the regularities of mind that Hume catalogues in his "science of man"; national characters spring from the customs and practices of local cultures;[9] individual characters are the character traits that are peculiar to a particular person, that capture his or her distinctive manners of behavior.

Thus, when evaluating a person's behavior, say, her giving money to the beggar she passes on the street, we must first see whether the action in question is connected to her in any significant way. Is she just doing what our common human nature would cause anyone to do? Is this just an

---

[8] Hume generally uses 'character' and 'mental quality' as synonyms for 'character trait'. Hume's reference to a person's disposition here points to the fact that we have to take into account the current state of mind of a person before we decide how her behavior counts as evidence for her character traits. On this point, see Brandt (1970, p. 35).

[9] I have discussed Hume's treatment of national characters in (1995). I argue there that national characters are derived from individual characters.

instance, say, of her wanting to help out the people she loves (a tendency Hume claims is part of the "original constitution of the mind" [T 2.2.6.6, SBN 368])? This could be the case if there were some preexisting relationship between her and the beggar; but otherwise, given Hume's plausible denial that human nature includes a general love of humanity (T 3.2.1.11, SBN 481), we would have to look for another explanation of her deed. Is she just doing what anyone of her culture would do? Probably not, but if it were customary among her people to give alms to all or most beggars one happens to encounter, the credit for her behavior properly would belong not to her *tout court*, but to her in so far as she exemplifies her national culture.

Is she doing something which indicates a durable disposition in her, a trait of character, something which speaks to who *she* is? This is the hard question that we will consider in what follows. For her giving could indicate her tendency to help others in need, in which case it is a manifestation of generosity, and thus something that we should admire her for. Or it could be that she gave to this beggar only because she was thought it would make her look good to others – her giving might then have its origin in her vanity, in which case we would disapprove of her giving in that it springs from a vicious character trait (T 3.3.2.10, SBN 597-598). Or it could be that she gave only because the beggar in question reminded her of a beloved aunt. The giving in this case might indicate only a vague tendency towards family-feeling. The problem, of course, is how we are able to distinguish between these different cases. How can we infer the trait which lies behind the observed act of giving?

It might seem that to answer this question we need only consider the intention from which the woman in question acted when giving to the beggar, for this will reveal to us the appropriate description of it. But four problems arise here. First, though an appeal to the woman's intention might help us in the case at hand, there are many other cases where someone's behavior manifests a trait even when it is clearly unintentional – consider the grimace of the impatient person held up by others' incompetence.

Second, even if the behavior is intentional, it is not always easy to discern the intention with which someone acts; there is "great difficulty of deciding concerning the actions and resolutions of men" (T 2.3.3.10, SBN 418).[10] But we usually can assume that certain kinds of the behavior

[10] Some interpreters even think that Hume has a problem accounting for our beliefs that others even have minds at all (see Pitson 1996). But Hume seems to suppose that we have a natural tendency to assume, by means of "presensations" (T 2.2.1.9, SBN 332), that other humans, not only have minds, but have ones which are organized more-or-less like

that we witness are the products of the will of the person in question, even if we cannot be certain as to how she sees the situation. In other cases, such as that of the grimacing impatient person, we assume that his behavior is caused by his passions. (Hume, of course, argues that a proper *philosophical* account of the will shows it too to be dominated by the passions, but he acknowledges that in common life people fail to grasp this point [T 2.3.3.8, 2.3.4.1, SBN 417, 419].) And in still others, say when we witness someone suffering from what appears to be an epileptic fit, we will attribute his behavior to his body and its chemistry, instead of to any features of his mind.

Third, supposing we can recognize the relevant intention, it is not often that the description under which the action is done sheds much light on its trait sources. In the case of the woman's giving, the relevant intention is likely to be nothing other than that of giving to the beggar. She rarely does it *intending* it to be a case of helping the needy, showing off in front of friends, or manifesting allegiance to family. As Bernard Williams notes, "the benevolent or kindhearted person does benevolent things, but does them under other descriptions [. . .]. The description of the virtue [or trait] is not itself the description that appears" in the agent's intention (1985, p. 10).[11] This point is even more obvious when vicious traits are considered; the unkind person, for example, rarely takes his actions to manifest unkindness. And this means that appealing to the agent's intention often does not help us in locating the character trait that the action in question manifests.

Fourth, even when an appeal to the agent's intention does lead us in the direction of a character trait, e.g. when the woman *does* intend her giving to the beggar to be a case of helping the needy, there remains the question of explaining what lies behind the intention. "An intention shews certain qualities, which remain after the action is preform'd" (T 2.2.3.4, SBN 349) since there must be *something* in her that caused her to form that intention on that occasion. But what exactly? For it could well be that she had the plight of the homeless on her mind that day only because she attended a public lecture on that topic the previous evening.

---

ours. Hume seems to have borrowed the term 'presensation' from Shaftesbury, who in his "Moralists" says that animals have "pre-sensations" of such things as what preparations to make when pregnant; humans have such pre-sensations "not in any proportionable degree." He goes on to suggest that our recognition of beauty is dependent on the same kind of pre-sensation "of a higher degree" (1900, pp. 76, 136).

[11] Hume subscribes to a version of this point when he argues for the distinction between the natural and the artificial virtues (T 3.2.1.4, SBN 478); the natural virtues involve behavior performed without the reference to the virtue in question, while the artificial virtues involve behavior that does refer to the relevant artifice (T 3.2.2.10, SBN 490).

And if she went to the lecture only to spite her friend who told her that she was too materialistic to care for others, the character trait that is responsible for her forming the intention to help the needy in this case might be spitefulness, not generosity. The point is that the intentions with which someone acts, just like the behavior she exhibits, are transitory events, and in order to connect them to the *person* in such a way that *she* gets credit or blamed for them, we must first find the durable dispositions which they manifest.[12]

It is clear, however, that Hume thinks that there are some such traits involved in every intentional act, and this means that the person in question is responsible for it, though *how* that responsibility is to be understood remains a problem. Consider, for example, a stingy person who intentionally does a good deed, the very same thing that a generous person would have done in the circumstances. Hume thinks that the *action* would count as good, since "every particular act of generosity, or relief of the industrious and indigent, is beneficial; and is beneficial to a particular person, who is not undeserving of it" (T 3.3.1.13, SBN 580). This does not mean that we would praise the stingy person for *generosity*; rather we would revise our understanding of his stinginess in light of his having done one laudably beneficial act. And it is in that revision that his responsibility for the good deed becomes apparent; it changes the way we understand him, at least to some extent. A similar point applies to most of us who are neither generous nor stingy. We do a sufficient number of things that are similar to what the generous would do so as not to qualify as stingy; but neither do our helping acts fall into a pattern of behavior – actions, reactions, feelings, and beliefs – that would mean we were generous. We are responsible for our beneficial acts in that others should take them into account in the assessment of our traits.

2. McIntyre's Suggestion

But we are left with the problem of how, from the behavior that we observe, we are able to discern the character trait that is its source.

[12] It is at this point that the Kant distinguishes between the purpose of the action and the motive of the action (the maxim from which it is done) (1993, p. 12 [Akademie edition, p. 399]). Whereas Kant's practical-reason based approach to morality leads him to say that a person's will – ultimately the person himself (p. 58, [Ak., p. 458]) – is what he must take ultimately to ground his maxims, Hume's observational perspective on behavior leads him to ask what it is about the person that leads him to will when he does, one way or another.

Hume, in his treatment of the will, suggests that this is a matter of causal inference, a suggestion which McIntyre uses as the starting point for her discussion (McI 196). Since Hume thinks that causal reasoning depends on prior observation of the constant conjunction of two types of events, a first step for explaining our capacity to recognize character traits requires us to go beyond our observation of a single action.[13] In the case of the woman and the beggar, we must observe her repeatedly giving to the needy (or some such thing) before we can infer that her giving here counts as generosity (and similarly for vanity or family allegiance). But even with this expanded observation of her behavior, it is still not clear how this causal inference is supposed to go. As we saw in §1, even if each piece of behavior can be traced to *some* mental event, we are often unable to determine much about it. How does this generic identification of a sequence of mental causes amount to the ascription of a trait, especially given the fact that we take a trait to be a feature of a person even in the periods between the emergence of the elements of the sequence (a generous person is still generous while she attends to her own needs)?

McIntyre suggests this problem can be solved if we see a character trait as a durable mental disposition or power. She acknowledges that Hume's analysis of causation includes a rejection of the coherence of causal powers (T 1.3.14.34, SBN 171) – they are mere hypostatizations of our associative tendencies – but she points out that Hume, in his discussion of the passions, is willing to countenance them despite these qualms. The miser, for example, takes pleasure in his money solely because it gives him the *power* to use it if he so desired. His knowledge of the social world allows him to make a causal claim about his riches even while he does not and will not actually use his riches in the standard manner (T 2.1.10.4-9, SBN 311-314). McIntyre suggests that a character trait is a similar kind of power. Observation of someone's performing a particular kind of passion-influenced or willed behavior in certain kinds of circumstances allows us to infer that she has the power to behave in that manner in those circumstances, even while those circumstances do not obtain (McI 199).[14]

---

[13] Hume does allow for the possibility of single-event-based causal inferences when we have developed a sense of how nature tends to work (T 1.3.12.3, SBN 131), but since we know that, in the case of human behavior, the same event might be the outcome of many different traits, we require more evidence than a single action for an inference to a trait.

[14] McIntyre rightly goes on to note, however, that, if this were the whole story, Hume would not be entitled to the realist assumptions he often makes about character. For how can such a power be the *cause* of a person's behavior if all the power amounts to is a counterfactual belief on the part of various observers? Hume's first definition of cause,

McIntyre argues, moreover, that Hume's treatment of the self supports this conception of traits. For even though the self is a bundle of perceptions, included amongst them are the various passions that are the primary causes of actions (T 2.3.3.4, SBN 415). Passions can be temporally extended (T 2.3.9.12, SBN 440-441), and when someone becomes accustomed to having a particular kind of passion play a dominant role in her mental economy, when it has become a "settled principle of action," that passion will no longer create a "sensible agitation" in her soul (T 2.3.4.1, SBN 419). And hence McIntyre concludes that Hume means to *identify* a person's character traits with the passions that tend to persist in her – or perhaps with several passions that persist in their relation to one another, as in strength of mind (McI 201; T 2.3.3.10, SBN 418) . That is why Hume sometimes uses the same terms to describe passions and traits: ambition (T 2.1.1.4, 3.3.2.13, SBN 276, 599), generosity (T 2.1.1.4, 3.3.1.11, SBN 277, 578), pride (T 2.1.1.4, 3.3.2.1, SBN 276, 592), benevolence (T 2.2.6.3, 3.3.3.3, SBN 367, 603) and others.[15]

Once traits are understood as durable passions any problems that might seem to arise from treating them as causal powers dissolve. For not only do we *infer* that someone has a causal power, what Hume calls *sympathy* – the tendency we have to become infected by the emotions of those around us – leads us actually to *feel* the relevant passions in him (T 2.1.11).[16] And this means that the character exists not merely as a projection from our causal beliefs, but as the passions we feel within him. For McIntyre, this combination of an inference to a causal power and the sympathetic-communication of passions is enough for Hume to sustain a kind of realism about characters.

---

after all, requires a constant conjunction of events of the cause-type and events of the effect-type (T 1.3.14.31, SBN 170). But here the purported cause, the character, is nothing but an observer's belief about a person hypostatized into a power. Is this any different from saying that a dormative power makes someone sleep or a motive power makes someone walk?

[15] McIntyre notes that, in the first *Enquiry*, Hume describes ambition, avarice, self-love, vanity, friendship, generosity, and public spirit as passions, while each of them is also a character trait that counts as a virtue or a vice (McI 200); see Hume (1975, p. 83, edited by L.A. Selby-Bigge and P.H. Nidditch) or (2000b, Section 8, Part 1, §7, edited by Tom Beauchamp).

Note that McIntyre's exegesis of Hume leaves his position quite close to that of Brandt, who takes traits to be intrinsic wants or aversions (1970).

[16] There are, of course, numerous problems afflicting Hume's treatment of sympathy. I have outlined an interpretation of it in which many, but not all, of these problems are avoided in (2005).

3. Four Problems

McIntyre's solution, however, is only a start, for the conception of character traits that she offers to Hume faces four problems. I will present them in increasing order of severity.

3.1. *Not All Character Traits Are Passion Complexes*

McIntyre's suggestion that traits are stable passion-complexes does not seem to account for all of the character traits that Hume acknowledges.

On the one hand, he is willing to identify traits, such as wit (T 3.3.1.27, SBN 590), patience (T 3.3.4.7, SBN 610), industry (T 3.3.1.24, SBN 587), and the like, for which he does not introduce parallel passions. Also, there are some traits that are culturally specific; racism, for instance, is only possible in a multi-racial society. Would Hume say that there is a passion of racism that only happens to emerge in such societies? Moreover, given Hume's methodological scruples, it seems unlikely that he would want to equate each trait with a specific passion; he is, after all, generally reluctant to multiply the number of primitive qualities that he will attribute to human nature (T 2.1.3.7, 3.1.2.6, SBN 282, 473).

On the other hand, Hume also recognizes character traits, such as prudence (T 3.3.1.24, SBN 587) or credulity (T 1.3.9.12, SBN 112-113), that include not only a passion-derived disposition to *behave* in certain ways, but also a disposition to *believe* certain things in a certain manner – namely, in this case, to believe either the causal truths connecting available means to one's ends, or the testimony of others. And once we recognize that such cognitive dispositions also contribute to character traits, it seems likely to suppose that Hume means them to play a role in those passion-involving characters that give McIntyre her strongest support. A generous person, not only acts givingly, but also *believes* that people in need merit support; the vain person, not only looks in the mirror more frequently than most, but also believes that he is better (in various respects) than those who surround him.

3.2. *Heterogeneous and Single-Act Character Traits*

Some character traits involve the repetition of a similar kind of behavior. A cheerful person, for example, smiles and laughs more frequently than most. But the behavior involved in other kinds of character traits does not involve this kind of homogeneity. The generous person gives money to the poor, helps a friend plant her garden, hosts a dinner party for an

acquaintance who is coming through town, shovels his neighbor's walk after a blizzard. The point is that the cheerful person's behavior can be identified as cheerful *prior to* his being identified as cheerful. And this means that we can, as McIntyre suggests, project back from his smiles and laughter to the character trait of cheerfulness; the persistence of one passion, or one kind of passion, might indeed underlie the trait. But the heterogeneity of the generous person's behavior makes the inference to the character trait more difficult. No one piece of the behavior she displays counts as generous until it is brought into connection with other pieces of her behavior. Someone might give to a beggar, shovel her neighbor's walk, or host a dinner party for an out-of-town guest without being generous; for she might not do these activities in the right spirit, as it were, or they might not cohere with the rest of her behavior in the way needed for them to indicate generosity. Thus our recognizing the generous person's patterns of behavior as instantiating generosity is *posterior to* (perhaps simultaneous with) our seeing her as generous.[17] This leaves it hard to see how the ascription of heterogenous traits can be treated as a causal inference from someone's behavior to the trait. And (despite Hume's anomalous comment about generosity at T 2.1.1.4, SBN 277)[18] it is hard to believe that just one persisting passion (or the reliable re-emergence of one kind of passion) is involved in the heterogenous behavior involved in cases like that of generosity.

Those traits that do not involve the *repetition* of behavior pose a similar problem. The perfidious person, for example, need only perform one grand betrayal once in order to merit that trait. The heroic might similarly not involve a *pattern* of activity, in that someone who rushes into icy waters in order to save a child from drowning need never again act in such a dramatic fashion in order to be properly called heroic (indeed, she might never again find herself in circumstances that allow for such action). These single-act-based traits seem not to require the *persistence* of any passions, and thus do not fit with McIntyre's model.

---

[17] G. Von Wright also makes this point: "[T]he results of all courageously performed acts need not have any 'outward' feature in common. Killing a tiger and jumping into cold water can both be acts of courage, though 'outwardly' most dissimilar. No list of achievements could possibly exhaust the range of results in courageous action [. . .]. [Moreover] the result of any courageous act could also have been achieved through action which was not courageous. Not even to have killed a tiger is a sure proof that a man is courageous. What is true of courage in the said respect is also true of the other virtues" (1993, p. 141).

[18] Even though, as I noted above, Hume does call generosity as a passion (T 2.1.1.4, SBN 277), he never gives any kind of sustained analysis of it.

3.3. *The "Depth" of Character Traits*

Nor does McIntyre's account explain the difference (in the non-single-act cases that she considers) between *mere* regularities in a person's behavior and the kind of consistency in action that points to a character trait. Someone might drink an average of two liters of water per day, but this does not mean that she has a character trait of excessive hydration. Or a man might always put his pants on with the left leg first, but this does not indicate a character trait. The point is that only some kinds of regularities of behavior have the kind of *depth* to make them indicative of a trait. For the very notion of *character* traits indicates that these regularities speak to *who someone is*; they *define* their bearers as particular kinds of persons; they are "connected with [their bearer's] being and existence" (T 2.1.8.8, SBN 302). Other, more superficial regularities, in contrast, are "in a manner separated from" (T 2.1.8.8, SBN 302) their bearer in that this person would remain who she was even if these regularities evaporated. (The parallel problem in the single-act cases is how we can distinguish between a *mere* doing, say, taking the dog for a walk, and the kind of act – saving the drowning child – that entitles one to a special designation as being of a special type such as the heroic.)

Recall that Hume says that before someone can be held responsible for an action it must be "infixed upon" him (T 2.3.2.6, SBN 411) so as to be connected to him as a *person*. Connecting an action to a character trait is a necessary condition for such an infixing. But is it sufficient? If we accept McIntyre's treatment of character traits, it is hard to see that it is. For her construal of traits leaves them as mere facts that describe someone, no different from such facts as that the person in question has a cardio-pulmonary system, that he ate a sandwich for lunch on January 7, 1987, or that he has never traveled south of the Equator. But these facts do not speak to who *he* is; they are incidental to him. Thus we must supplement McIntyre's account with an explanation of how a regularity in behavior comes to have the depth that makes it a *character* trait, of how such behavior is connected with the *person.*

3.4. *The Normativity of Character Traits*

A related problem to 3.3 has to do with the normativity of Humean character traits. For at least some of them, the virtues and vices, are not only deep features of persons, they are *moral* features of them – they "command" our approbation or disapprobation (T 3.1.2.4, 3.3.1.13, 3.3.1.18, 3.3.1.25, 3.3.1.27, 3.3.5.1, SBN 472, 580, 584, 589, 590, 614). Hume even says that our moral language itself reflects this feature of

characters: "The distinction [between virtue and vice] [. . .] being so great and evident, language must soon be moulded upon it, and must invent a peculiar set of terms, in order to express those universal sentiments of censure or approbation".[19]

Note that this means that it is possible for someone to be generous, and thus to merit approval, even though others (and perhaps she herself) fail to discern this trait in her. By taking moral approval and disapproval to be built into trait recognition (at least for some traits), Hume seems to leave room for a kind of objectivity in his moral theory, despite his commitment to a subjectivist moral epistemology in which moral approval is a matter of feeling, not ratiocination.[20] On McIntyre's approach to trait recognition, however, where traits are discovered primarily by causal reasoning, it seems hard to see how Hume can moralize trait recognition without violating his subjectivist denial that moral "facts" can be discovered by causal reasoning.[21]

4. The Indirect Passions

In this section, I will start my attempt to resolve these four problems with a consideration of how a single-act character trait might come to have depth. Or, to put it another way, how is it that an ephemeral piece of behavior (the woman's saving of the drowning child) can come to define who someone is for her entire life? For there is a sense in which, when

---

[19] Hume (1975, p. 274, edited by L.A. Selby-Bigge and P.H. Nidditch) or (1999, Section 9, Part 1, §8, edited by Tom Beauchamp). Further references to this work will be made parenthetically as "EPM" followed by Section, Part, and paragraph numbers (as given in the Beauchamp edition), followed by "SBN" and the page number as given in the Selby-Bigge and Nidditch edition.

Also: "The people, who invented the word *charity*, and used it in good sense, inculcated more clearly and much more efficiently, the precept, *be charitable*, than any pretended legislator or prophet, who should insert such a *maxim* in his writings" (Hume 1987, p. 229); further references to this work will be made parenthetically as "Es" followed by the page number. See also T 3.3.1.16, SBN 582; EPM 9.1.6, SBN 272; Es 227. Recognizing someone's generosity, for example, and approving of it are inseparable because approval-worthiness is built into the very nature of the character trait of generosity.

[20] David Norton in his (1982) goes so far as to take Hume to be a moral realist, given his treatment of virtues and vices as traits demanding approval and disapproval.

[21] McIntyre does discuss what she thinks is "an important normative element in the concept of character itself," namely the relativity of our trait concepts to the standards set by human nature (McI 201) – a point which I discuss in §6.4 below. But this normativity is different from the moral normativity involved in virtuous traits' "commanding" us to approve of them.

someone is a hero, we not only take it to make a difference to who she is in the period following the act, but we also take it to be true of her retroactively; we think of her as having been incipiently heroic in the period prior to her manifestation of heroism. But how can Hume explain our capacity to think of people in this way?[22]

If we look only at Hume's treatment of persons and of the faculty of understanding in Book 1 of the *Treatise*, it seems that he has nothing to say in answer to this question. For if persons are, as he suggests, "bundles of perceptions" (T 1.4.6.4, SBN 252), how can the perceptions that amount to the salient single act (the saving of the drowning child) play a special role in defining who someone is? That those perceptions are part of her bundle is just one more fact about her – and as we saw in the discussion of depth, §3.3 above, it is hard to see how one can separate deep facts about someone from mere facts that are true of her. Her perceptions-from-saving-the-child are part of her bundle, but so are the perceptions arising from her eating lunch on January 7, 1987 or from her water-drinking habits. Hume, in Book 1, treats all the perceptions to be found within someone as on a par, and thus he will be unable to explain how some of those perceptions can be tied up with someone's character in socially meaningful ways. Nor will he be able to explain how we see certain features of a person as indicating something special about her, not just at a moment, but throughout her life.

But we need not restrict our view of Hume's treatment of persons to Book 1's "bundle" view. For when he presents that view he is quite clear that he has restricted his attention to the self "as it regards our thought or imagination," postponing his consideration of the self "as it regards [. . .] the concern we take in ourselves" until he has examined the passions (T 1.4.6.5, SBN 253). And when he considers the passions in Book 2 of the *Treatise* the bulk of his attention is devoted to the person-involving "indirect passions" of pride and humility, and love and hatred, where the former pair are directed towards the self and the latter two are directed towards another. We need not concern ourselves here with the specific details of Hume's associationist mechanism for these passions, the "double relation of ideas and impressions." Generally, what happens is that when we recognize some fact about a person (either ourselves or another) that engages us emotionally, positively or negatively, we transfer that positive or negative feeling from the fact to the person. This emotion, with its attendant focus on the person, is, as the case may be, pride, humility, love, or hatred. In feeling proud of our house, for

[22] I have discussed a more general version of this question in Ainslie (1999). The answer I give here and in the section that follows borrows form the argument of this other paper.

example, our positive assessment of our house is refocused into a positive feeling about ourselves, in so far as we own the house (T 2.1.5). Note that Hume treats these passions in a very generic fashion in that he openly declares his lack of interest in spelling out the differences between various kinds of positive or negative feelings towards persons (T 2.3.9.31, SBN 448). His general point is that the indirect passions convert *a theoretical belief* about the facts that are true of someone into *a practical attitude* towards the person such that those facts are seen to suffuse who she or he is.

We can now see how Hume can explain how we ascribe single-act character traits to people. We feel the passion of love towards the woman who saves the child because her saving of the child causes us pleasure (and we need not have a special connection to the child in order to care about his being saved; sympathy conveys to us the feelings of gratitude of the child himself and of his family and friends). And love converts our feeling of pleasure at her act into a feeling of pleasure at her *overall*, so we now characterize her in terms of this act; "we can never think of [her] without reflecting on these qualities" (T 2.2.3.4, SBN 349) in her that are responsible for her behavior.

A similar story would be available in the case of McIntyre's version of causal powers. In order for these to be more than mere facts about someone, in order for them to speak to who the person in question *is*, we must love or hate him for them (where the causal powers are our own, we must feel pride or humility in light of them). And this passional response will integrate these powers into our conception of him. The difference between mere regularities in behavior and a character trait is that the trait causes an indirect passion in us because of its salience to us. In particular, those traits that are virtues cause us pleasure because of their being useful or agreeable to their bearers or to those around him (T 3.3.1.30, SBN 591); and thus "we approve of his character, and love his person" (T 3.3.3.2, SBN 602) when we recognize his virtue (traits that are vices cause pain because of their being detrimental or disagreeable to their bearers or to those around him; they lead us to hate him for his vice).

5. General Rules and Maxims

The argument of the previous section has gone only a small way towards addressing the four problems raised in §3. While our finding depth in at least some characters (single-act and homogeneous ones) has been

accounted for, we are still owed an explanation of our capacity to recognize heterogeneous characters (including those that involve beliefs as well as passions) and of the normativity of character ascriptions. My route into solving these problems will be the issue of normativity. For, despite his giving in the indirect passions a merely causal explanation of our attributing traits to persons, Hume seems to think that there are occasions where we *should* attribute such traits. As we saw in §3.4, virtuous character traits *command* our approval, where approval of the traits and recognition of them go hand in hand. But if someone is not in a position where the causal conditions needed for the indirect passions to operate in him are satisfied, how can Hume say that he ought to recognize the trait?

It might seem that in the case of virtues and vices Hume answers this question when he requires that we make certain corrections to our sentimental responses to one another in order to take up a "moral point of view" (T 3.3.1.15, SBN 581-582). But this cannot be the whole story. First, Hume leaves this requirement underexplained. He says that we must follow it in order to be able to communicate amongst ourselves, but this does not seem true, in that we can communicate with one another using the language of self-interest without much difficulty (see Sayre-McCord 1994, especially p. 215). Second, Hume seems to introduce the normative element into character trait recognition well before he talks about the corrections required for moral evaluation. For, when describing the indirect passions, he says:

> ['T]is evident, that if a person full grown, and of the same nature with ourselves, were on a sudden transported into our world, he wou'd be very much embarrass'd with every object, and wou'd not readily find what degree of love or hatred, pride or humility, or any other passion he ought to attribute to it [. . .]. [A]s custom and practice have brought to light all these principles, and have settled the just value of every thing; this must certainly contribute to the easy production of the passions, and guide us, by means of general establish'd maxims, in the proportions we ought to observe in preferring one object to another. (T 2.1.6.9, SBN 293-294)

This suggests that Hume thinks that custom, practice, or what he calls in the paragraph preceding this one, "general rules" (T 2.1.6.8, SBN 293), are what transform the indirect passions from merely causal responses that some people happen to feel into attitudes that we ought to take up towards one another. Hume continues the passage by saying that his introducing these general rules will serve to "obviate difficulties, that may arise concerning" how some causes of the indirect passions "operate so universally and certainly, as they are found to do" (T 2.1.6.9,

SBN 294). That is, he reaffirms that the problem of normativity that we are confronting in this section is supposed to be solved by these general rules. But how exactly?

Consider a homogeneous trait such as cheerfulness. Many people find cheerfulness agreeable, and because of this they feel the indirect passion of love towards people with this trait and consequently see them as cheerful *persons*. In this sense there is a custom or general rule linking this trait to positive person-oriented passions. Now, what of the person who recognizes someone's cheerfulness but who fails to care about it, the person who finds it no more interesting than a tendency to drink an average of one liter of water per day?[23] Why ought he to ascribe a character trait of cheerfulness, with the depth it entails, to the person in question? Hume's point seems to be that others' tendency to respond to this person in a similar way serves to *convert* his cheerfulness from a mere fact that is true of him into something that suffuses him as a *person* even for those who do not feel strongly about that fact; the general or customary ascription of that trait to him constructs a "rank" (T 2.1.6.8, SBN 293) into which this person falls even if some people fail to share the common feelings. For when people take his cheerfulness to matter, they treat him differently (by, say, volunteering to work with him, or inviting him along on social occasions), and when enough people treat him differently because of his cheerfulness, it will be a salient feature of him even for those who fail to see what makes good cheer so special. The social world will come to include cheerfulness as a meaningful social marker in such a way that someone who fails to acknowledge it will be "embarrass'd" – in the 18th-century sense of confused or at a loss – in her failing to grasp the established social distinctions, for she will misunderstand how the person in question fits in to her society.

What of the heterogeneous character traits such as generosity or, to use an example to which Hume devotes more attention, wittiness (T 2.1.7.7, SBN 297)? There clearly are "general rules" that apply to them, too, since, growing up in a society where we encounter generous and witty people, we are taught what kinds of behavior can be expected from them and that they make a difference to who the persons in question

[23] In addition to the case of someone who recognizes the cheerfulness but who fails to care about it, there is also the case of someone who misses out on the cheerfulness altogether. This raises a different kind of normative question: Why ought someone to recognize causal truths for which she has evidence? As this is a general issue within Hume's epistemology, I will not address it now. But note that my description, below, of the parallels between Hume's treatment of causal reasoning and of character-trait ascription indicates the kind of line he should take on this question.

are. The "general establish'd maxims" of wittiness, for example, might include such things as: "the witty person punctures the pretensions of the pompous," or "the witty person enlivens a dinner party." But this still leaves us with a question. Given that maxims like these pull together a heterogeneous range of behavior, how did they arise? Hume does not answer this question directly, but he seems to appeal to the historical shaping of our affective responses to one another. In a case like wit, we have come to see a capacity to entertain verbally as making a difference to who someone is because our primitive capacity to take pleasure in one another's utterances has been shaped and molded by the traditions we inhabit. So in one culture wit might involve appropriate allusions to literary classics, while in another it might involve puns and wordplay.

This might seem to introduce a kind of homogeneity into what I have been treating as heterogeneous traits, in that the maxims of wit group together certain kinds of behavior as typical for its bearers. But note that, unlike the action-descriptions involved in a homogeneous trait such as cheerfulness, the action-descriptions involved in these maxims are dependent on how others respond to the behavior in question, for someone counts as "puncturing pretensions" or "enlivening a party" only to the extent that people react to him in the relevant ways: they must find his conversational contributions pleasing. This will depend on his timing, his delivery, how his comments comport with the expectations people have of him, and so on. And so no one kind of comment or conversational contribution will satisfy the maxims. As Hume notes:

> No one has ever been able to tell what *wit* is, and to shew why such a system of thought must be receiv'd under that denomination, and such another rejected. 'Tis only by taste we can decide concerning it, nor are we possest of any other standard, upon which we can form a judgment of this kind. Now what is this *taste*, from which true and false wit in a manner receive their being, and without which no thought can have a title to either of these denominations? 'Tis plainly nothing but a sensation of pleasure from true wit, and of uneasiness from false, without our being able to tell the reasons of that pleasure or uneasiness. (T 2.1.7.7, SBN 297)

Our inability to "tell the reasons" for wit lies behind the heterogeneity of what counts as a manifestation of this trait, since whether a person's verbal output counts as a sign of his wittiness all depends on what we generally take pleasure in. It turns out, then, that the only kind of homogeneity connected with this heterogeneous trait is that the behavior that manifests it elicits a common response from most of those who encounter it. A similar story could be told about generosity or other

heterogeneous traits. They are the product of the evolution of feeling in our culture as certain kinds of behavior have come to be seen to constitute a kind because of their tendency to elicit from us common feelings of pleasure or displeasure. They "in a manner receive their being" from our shared, historically shaped, sensibilities.[24]

If Humean character traits are, as I have argued, the offspring of our affective responses (the indirect passions) to one another as regularized by "general rules," it might seem that the kind of normativity that is involved in character trait ascription is somewhat thin, having more to do with social conformity than with any more robust kind of *ought*; for these rules only specify the way our particular society tends to make sense of behavior. And why must we go along with *that*?

Hume's answer seems to be that we do not have much choice in the matter. If, according to the general rules prevalent in a society, a certain pattern of behavior indicates a trait, then someone who manifests the behavior has the trait, and has it in a way that makes him who he is, whether he likes it or not. A racist, for instance, might vehemently deny that she harbors discriminatory attitudes, but if others observe them in her behavior, she merits that trait despite her disclaimers. This is not to say that Hume thinks that we should blindly accept whatever trait-concepts happen to exist in our society. As he himself demonstrates in the case of the "monkish" virtues of "celibacy, fasting, penance, mortification, self-denial, humility, silence, solitude" (EPM 9.1.3, SBN 270), we can step back from social customs to see if we *should* continue to use them to make sense of one another. In Hume's case, the relevant standards for such an evaluation are functional ones: does the trait isolate a kind of behavior that is useful or detrimental, agreeable or unpleasant to the trait's bearer or to those who surround him or her? That is, Hume takes the standards we are to use when deciding whether a particular trait concept should be abandoned or retained to spring from the very features of our passions that are relevant to trait construction. Since the indirect passions at the root of our trait-ascriptions depend on our finding something pleasing or displeasing in one another's behavior, we must use this aspect of our passions when evaluating our society's moral customs.

---

[24] This leaves Hume's treatment of character traits very close to David Wiggins' cognitivist treatment of values as the historically shaped offspring of our affective sensibilities in (1987). But note that Wiggins takes himself to be offering a *corrective* to what he sees as Hume's overly subjectivist account of value. Given that Wiggins focus there is "Of the Standard of Taste" (Es 226-249) it is not surprising that he overlooks the *Treatise* account of the interplay between general rules and the indirect passions.

It will be useful at this point to compare the kind of normativity that emerges in this Humean treatment of character trait ascription with the kind that emerges in his treatment of causation. Hume himself hints that these two issues are similar when he says that the associationist mechanism he gives for the indirect passions has "a great analogy" with the associationist mechanism he gives in Book 1 to account for our beliefs about causation (T 2.1.5.11, SBN 289-290). Recall that, in that case, Hume offers such a mechanism because he thinks that there is no other way to account for our recognition of causal connections, given that there are no real "necessary connexions" waiting to be discovered by reason (T 1.3.6). We can see a similar rationale behind Hume's introduction of the indirect passions. Since he is unable to explain how reason can discern which of the facts that are true of a person define her as who she is, he offers an alternative associationist story for our forming these beliefs. Just as human nature comes equipped with regularity-oriented association-engendered feelings that lead us to view the world as causally structured, so also it comes equipped with person-oriented association-engendered feelings that lead us to construct a social world in which only some of persons' features make them into who they are.

Moreover, just as we have seen that Hume relies on "general rules" in his account of character-trait ascription in order to introduce into it an element of normativity, so also he relies on "general rules" in his account of causation in order to introduce into it a similar kind of normativity (T 1.3.13.7-12, SBN 146-150; T 1.3.15) – indeed, he openly models his passional general rules on the causal ones (T 2.1.6.8, SBN 293).[25] The causal-belief-forming associationist mechanism responds to experienced regularities of conjunctions of types of events. But since we might happen to have *experienced* two types of events as conjoined even though they generally *occur* independently of one another, there are cases where we *ought* not to form a causal belief even though the causal preconditions for the associationist mechanism are satisfied in our particular case. The normativity here arises because were we to have experienced things differently or more extensively, our causal belief would have been undermined by what actual conjunctions of events occur in the world.

[25] Hume says that the influence of general rules on the passions "may be accounted for from the same principles, that explain'd the influence of general rules on the understanding. Custom readily carries us beyond the just bounds in our passions, as well as in our reasonings" (T 2.1.6.8, SBN 293). Note that despite Hume's seemingly negative assessment of the influence of general rules in this quotation, he goes on to make the claim about their role in constructing a normative social order in the paragraph that follows it, quoted above (p. 94). Hume displays a similar kind of studied ambivalence in his treatment of causal general rules (T 1.3.13.11-12, SBN 149-150).

The causal mechanism for our causal beliefs sets standards which we can use to judge the causal beliefs we happen to have formed, just as the causal mechanism for our character-trait ascriptions sets standards which we can use to judge the trait ascriptions we happen to have felt.

There is a notable difference between the two kinds of normativity here. In the causal case, the standards we use to judge causal beliefs are the actual events in the world (as in the first definition of cause that Hume gives at T 1.3.14.31, SBN 170), events which are independent from our associative propensities (T 1.3.14.28, SBN 168). But the standards we use to judge character-trait ascriptions relate to the social world we live in, a world which is itself constructed from our associative tendencies. Since someone *is* cheerful or generous if, in her society, the customary passional response to her indicates that these qualities make a difference to who she is, we *ought* to recognize them when interacting with her. And, as we saw above, this means that not only can we challenge a particular character-trait ascription when it fails to pin down a person's social place, we can also challenge the very existence of the social place itself. This second kind of challenge has no analogue in the case of causal beliefs, for whether or not there is a constant conjunction of two types of events is not open to revision by us.

Moreover, when it comes to the traits that are virtues or vices, there are two different ways in which we can challenge their existence as meaningful social markers. Recall that Hume thinks that virtues are those traits that are useful or agreeable to their possessors or to those in their circle. This means that, on the one hand, a moral dissident can argue that the behavior specified by a trait fails to promote the relevant interests. As Hume puts it: "[E]rroneous [moral] conclusions can be corrected by sounder reasoning and larger experience,"[26] in that claims about the useful are causal claims having as much objectivity as any Humean causal claim. If the dissident can show, say, that self-denial does not serve to improve our lots, but in fact only "stupifies the understanding and hardens the heart, obscures the fancy and sours the temper" (EPM 9.1.3, SBN 270) – consequences widely admitted to be undesirable – then she will have the means to vindicate her heterodox rejection of this trait as approval worthy. On the other hand, when it comes to traits the approval of which depends on their agreeableness, Hume's subjectivism looms larger. When someone just does not like, say, what others in his society take to be witty, then he simply stands as an outsider. Indeed his deafness to their sense of humour might leave him unable to identify the

[26] "A Dialogue," published in conjunction with both the Selby-Bigge/Nidditch (p. 336) and Beauchamp (§36) editions of the second *Enquiry*.

bearers of this trait at all, except by indirect signs, such as others' laughing at their remarks.

But both kinds of challenge to the prevalent practices of trait ascription have a rather tenuous status, for someone's dissent from the customary ways of making sense of behavior is not by itself sufficient to undermine that trait as a social marker. That would require a wider renegotiation of the social world itself. For if we find that our customs of trait ascription have gone off the rails in some respect, as Hume felt had happened with the monkish virtues, our reflective verdict has only a slight and indirect impact on the relevant "general rules" of character ascription. Even after Hume had debunked humility and the like, most Europeans continued to take these qualities to be virtues; it was only after secularization had penetrated throughout society that these virtues were finally moved to the column of the vices, as Hume had recommended. In the meantime, the social world will continue to include the impugned traits, and the moral dissident who challenges them will continue to find his life defined in terms of them, whether he likes it or not. The most he can do, as Hume himself did, is to try to ignore his being defined by them: "I see not what bad consequences follow, in the present age, from the character of an infidel; especially if a man's conduct be in other respects irreproachable" (Greig 1932, p. 106).

The Humean approach to character traits that I have sketched here treats traits as precipitates from the indirect-passional responses that are common in a society. I have suggested that seeing traits in these terms avoids the problems that McIntyre's interpretation faced with the heterogeneity of some traits (problem 3.2), the depth of traits (problem 3.3), and the normativity of traits (problem 3.4).

Note also that this treatment does not require us to treat traits as passion-complexes that persist within their bearers. For so long as we take the various actions that are supposed to be manifestations of the trait to have mental sources (they are not mere bodily happenings like the epileptic's seizures), it does not matter that we cannot pin down exactly which passions and beliefs are responsible for them. After all, Hume uses the generic term 'mental quality' as a synonym for character trait (see n. 8). It is also possible for the maxims of a trait to have reference to beliefs and other kinds of attitudes; we no longer need to treat traits as always being directly bound up with action. This means that the interpretation I have offered here also avoids problem 3.1.

6. Conclusions

The approach to trait ascription that I have examined here is intimately linked to the following six elements of Hume's moral theory.

6.1. *All Character Traits Are Virtues or Vices*

Up to this point in my discussion, I have kept separate the issue of how we identify a person's character traits from the issue of how we decide which of those traits are virtues or vices. Common wisdom in contemporary moral theory includes a strong distinction between moral traits, such as benevolence or vanity, and non-moral ones, such as wittiness or gullibility (see, e.g., Brandt 1970, p. 23). Hume, however, denies that any such distinction can be made (although he does admit that we can distinguish between behavior that will be influenced by rewards and punishment, and behavior that is immune from such external control) (T 3.3.4, EPM App.4). We can now see why Hume takes this line. Since the indirect passions are involved in character trait ascription, and since their structure requires that their causes be pleasant or unpleasant, we must find the behavior that is to count as exemplifying a character trait to *matter* to us, one way or the other, by our being pleased or displeased by it, before we can ascribe a trait to the person in question. In so far as we consider the behavior "in general, without reference to our particular interest" (T 3.1.2.4, SBN 472), this pleasure or displeasure qualifies as moral approval or disapproval of the trait, and thus the trait qualifies as a virtue or vice.

Hume's point is that the way we consider one another as bearers of traits, when we consider one another impartially, *is* the way we make sense of one another morally. A moral perspective on one another need not commit us to metaphysical conceits such as God-the-judge; morality just is a label for our way of making sense of one another when we take an interest in the kinds of persons we are. To be more exact, since our moral concern is oriented towards one another's thought and behavior, the moral perspective focuses on the kinds of persons we are in so far as they speak to features of our minds – hence the interest in character traits as mental qualities.[27] And thus the part of morality that Hume considers "natural" (T 3.3) is not primarily a set of constraints on our behavior that we must recognize when we deliberate; it is instead the outcome of a

[27] "To suppose measures of approbation and blame, different from the human, confounds every thing. Whence do we learn, that there is such a thing as moral distinctions but from our own sentiments?" (Es 595).

perspective we can take on our own and others' behavior, when we examine it to see what it says about who the person in question is. Hume summarizes his conception of morality by saying that the "ultimate test of virtue and merit" is to see "if there be no relation of life, in which I cou'd not wish to stand to a particular person [. . .]. If he be as little wanting to himself as to others, his character is entirely perfect" (T 3.3.3.9, SBN 606).[28]

One consequence of Hume's broad conception of virtue is that he has to admit what might be called "traits of taste" into his catalogue of the virtues, for we do categorize one another as such things as film buffs, fitness freaks, the outdoorsy, or intellectuals, on the basis of the activities we go in for. Hume himself frequently invokes these traits of taste in his *History of England*, when summarizing the characters of the various kings and queens: Henry I, for example, was "addicted to women" and "hunting was one of his favourite amusements"; James I had a "rustic contempt of the fair sex," while his wife, Anne of Denmark, "loved shows and expensive amusements, but possessed little taste in her pleasures" (Hume 1983, Vol. 1, Ch. 6, p. 277; Vol. 5, Ch. 49, pp. 122, 221). Moreover, Hume's "test of virtue and merit" requires the inclusion of traits of taste. I could not wish to stand in the relation of husband to someone who was outdoorsy or a fitness freak, and so these traits are to that extent vicious. Of course, the outdoorsy fitness freak might feel the same way about an intellectual film buff, but this only indicates that Hume's conception of virtue includes some latitude when it comes to traits of taste. Different people find different kinds of activities agreeable and, although what we enjoy makes an enormous difference to whom we want to spend our time with, we also allow that others will have their own distinctive tastes. As Hume says when discussing literary works, "[w]e choose our favourite author as we do a friend, from a conformity of humour and disposition [. . .]. Such preferences are innocent and unavoidable, and can never reasonably be the object of dispute, because there is no standard, by which they can be decided" (Es 244). The parallel point for traits of taste means that each person will have a conception of perfect virtue inflected by her or his preferences, even while there is broad overlap in the main set of traits he or she acknowledges.[29]

---

[28] See also the description of Cleanthes at EPM 9.1.2, SBN 269-270.

[29] Note that there is no conflict between Hume's admission of traits of taste and his requirement that traits are to be evaluated in abstraction from our self-interest. The point is that my self-interested desire for a film buff partner should not be allowed to distort my recognition that my date is outdoorsy.

6.2. *Moral Points of View*

I have noted that we must regard one another impartially if our character trait ascriptions are going to yield attributions of virtues and vices (T 3.1.2.4, SBN 472). Hume says that, in addition to this bracketing of our self-interest, we must also place ourselves in a "*steady* and *general* point of view" (T 3.3.1.15, SBN 581-582) – the view of the bearer of the trait herself or of those who are part of her "narrow circle" (T 3.3.3.2, SBN 602) or who have "an intercourse" (T 3.3.1.17, SBN 582), "commerce" (T 3.3.1.18, SBN 583), or "a connexion" (T 3.3.1.30, SBN 591) with her – to correct for the variation in our sentiments caused by the variations in our perspectives on one another's character. I suggested earlier (see p. 94) that Hume's explanation for why we must make this correction – namely, in order to increase the ease of communication – is inadequate. But it is now clear that Hume has a better explanation available to him. For we have seen that whether a person has a virtue or vice, whether he has an approval-worthy or disapproval-worthy trait, is a feature of the social world, even if it is not a fact discoverable by the understanding (neither by demonstrative nor by causal reasoning). The impartiality requirement indicates that if we are to pin down how the person's behavior locates him in the social world we must ignore any distorting influence created by our self-interest. We will not discover where *he* fits in if we are guided by where *we* would like him to fit in.

The second kind of correction has to do with the nature of the information that we must have access to if we are to capture someone's character traits accurately. Consider the woman giving to the beggar, discussed above. Someone must know her quite well in order to be able to tell whether her giving counts as generosity or vanity, for the giving must be brought into connection with a good portion of the rest of her behavior before it can cohere in a pattern indicative of a trait. Most of us, not being appropriately located to have access to this information, must defer to the evaluation of someone who is – supposing, that is, that his verdict satisfies the impartiality requirement. Of course, how well we must know someone in order to have access to the kind of information relevant to assessing her traits will depend on the trait in question. "Heroic virtue" such as courage or ambition (T 3.3.2.13, SBN 599) is better observed from a distance, as when civilians laud their generals. The virtues of the "good" such as gratitude or compassion (T 3.3.3.3, SBN 603) can only be discerned by those who are intimate with their bearers.[30]

[30] The correction Hume thinks is necessary for "virtue in rags" (T 3.3.1.19, SBN 584)

6.3. *Moral Evaluation and the Passions*

Hume's taking moral evaluation to involve an assessment of a person's traits of character explains why he repeatedly links the moral sentiments and the indirect passions. He says, for example, that "these two particulars are to be consider'd as equivalent, with regard to our mental qualities, *virtue* and the power of producing love or pride, *vice* and the power of producing humility or hatred" (T 3.3.1.3, SBN 575) and that the indirect passions are "perhaps the most considerable effect that virtue and vice have upon the human mind" (T 3.1.2.5, SBN 473; cf. also T 2.1.7.2, 2.1.9.1, 3.3.5.1, SBN 295, 303, 614). We now see that Hume *must* have the indirect passions play a role in moral assessment, because it is only by means of them that we can recognize character traits as defining features of one another.

I suggest that Hume's point here is not merely an artifact of his system, or his approach to moral philosophy, but a lasting insight. His recognition that there is a radical difference between seeing the world as merely a collection of events and seeing it as including *people*, some facts about whom are definitory, seems sound. The category of person is an irreducible feature of the moral world. This is not to say that one must follow Hume by explaining this feature in terms of association-engendered person-oriented passions. Kant, for example, takes the conception of oneself (and other rational beings) as persons to be part of what it is to be an autonomous agent. Hume's passion-derived treatment, however, has the advantage of explaining our moral attitudes when we attempt to evaluate ourselves and others, not just when we deliberate about what to do.

6.4. *The Rarity of Virtue and Vice*

Recall that Hume distinguishes between three different kinds of traits, those common to humans in general, those distinctive of a nation (or a cultural grouping of some other sort, such as a profession), and those "peculiar [. . .] to a particular person" (T 2.3.1.10, SBN 403; see p. 82, above). There is a sense then that any character trait (and thus any virtue

---

seems to be of a different sort from the others. It has less to do with describing the perspective for best recognizing characters than with what we ought to count as relevant for character attribution. Hume reminds us here that the patterns of activities springing from a character are constrained by the circumstances of the agent in question. His point seems to be that the general rules governing character attribution should leave open possibilities such as the generosity of a man in a dungeon; we ought to be sensitive to the nuances of behavior exhibited by people *in their circumstances* when trying to make sense of their characters.

or vice) must be somewhat rare. Since it points to something distinctive about its bearer's behavior, it would no longer indicate anything special about *her* if too many people shared it. As Hume says:

> [T]he honourable appellations of wise and virtuous, are not annexed to any particular degree of those qualities of *wisdom* and *virtue*; but arise altogether from the comparison we make between one man and another. When we find a man, who arrives at such a pitch of wisdom as is very uncommon, we pronounce him a wise man: So that to say, there are few wise men in the world, is really to say nothing; since it is only by their scarcity, that they merit that appellation. (Es 83; cf. Es 238)

The point is that we make our character trait attributions in light of our expectations about how people typically behave in various situations, "the ordinary course of nature in the constitution of the mind" (T 3.2.2.8, SBN 488; cf. T 3.2.6.9, 3.3.3.2, SBN 532, 602). Only when someone behaves differently in a notable way does she merit a trait ascription (T 3.2.1.18, SBN 483).[31]

Note that the fact that each virtue or vice is rare does not mean that only a few people have such traits. Instead, Hume's view is that each of us is properly described as bearing *many* (rare) virtues and vices, even if no single virtue or vice is common to all of us. As we have seen, Hume has an expansive conception of morality in that any trait in terms of which we make sense of one another's behavior counts as a virtue or vice. One person might be generous, dull, and unreliable, while another might be a witty tightwad: Hume is no unity-of-the-virtues theorist.[32] The character sketches in the *History of England* show us the complex ways in which Hume thinks that traits can be combined in different people (see Dees 1997).

---

[31] Annette Baier argues, in contrast, that Hume leaves open the possibility of everyone's having certain natural virtues, such as parental solicitude; morality no longer becomes a project of "selecting the few who are to pass through the narrow door into heaven" (1991, p. 208, see also pp. 199, 207-209). In fact, Hume says only that "natural affection" for children "is the duty of every parent" (T 3.2.1.5, SBN 478), where a duty is something the omission of which counts as a vice (T 3.2.5.4, SBN 517). Hume's point seems to be that since it is a part of the human nature to have concern for one's children, failure to be so moved has come to be recognized as a vicious trait. The rare occurrence of parental neglect is a vice, while the common occurrence of parental affection and care is a duty, not a virtue. (Hume also says that being an "indulgent father" [T 3.3.3.9, SBN 606] counts as a virtue, but there is no sign that Hume expects this trait to be common.)

[32] Hume says of Charles I: "The character of this prince, as that of most men, if not all men, was mixed; but his virtues predominated extremely above his vices" (1983, Vol. 5, Ch. 59, p. 542).

Hume accounts for our being oriented towards features of persons that set them apart, that distinguish them from others, by incorporating a rarity requirement into his treatment of the indirect passions. Something can be a cause of this passion only if it is "peculiar to ourselves, or at least common to us with a few persons" (T 2.1.6.4, SBN 291). Once more we see that Hume explains why our social world of persons has the shape it does in terms of certain primitive features of our passional reactions to one another.

6.5. *No Change of Character*

In the course of arguing for the moral equivalence of talents and the more traditional virtues (see 6.1), Hume considers the suggestion that these traits are different in kind because talents are innate and thus immutable, while the traditional virtues result from voluntary choices. He responds that virtues are just as innate and immutable as talents, "it being almost impossible for the mind to change its character in any considerable article, or cure itself of a passionate or splenetic temper, when they are natural to it" (T 3.3.4.3, SBN 608). This is perhaps a shocking conclusion, for we might think that someone could undergo a moral conversion, say by leaving his stinginess behind for a new-found generosity. But Hume's point is what we should expect given the account of trait ascription explored above. As I argued in §4, Hume relies on the indirect passions to explain our capacity to see the world as containing not just events, but *persons*; these passions are an associative mechanism that have as an outcome our seeing someone as characterized by some of her features in such a way that they make a difference to who she is *throughout her whole life* (see T 2.1.6.7, SBN 293). This is why the heroic person is retroactively thought to have been incipiently heroic even before she did the heroic deed. And this is why Hume denies the possibility of character change. The associative mechanism involved in character trait ascription always leads to our thinking of the person as having had the trait throughout his lifetime.

This means that, for Hume, when the person who had been behaving stingily starts to be generous we say, not that he has *changed* traits, but that we were *wrong* all along in our ascription of stinginess to him. What we thought was a pattern of behavior indicative of stinginess turns out only to have been the early portion of a pattern of generous behavior.

6.6. *The Anomalous Place of Justice*

Hume tends to treat justice and the other artificial virtues as if they are just like the natural virtues – namely, character traits that command approval (T 3.2.2.23, 3.3.1.12, SBN 498, 579-580). But the two types of virtues are quite different if my interpretation of character traits is right. First, while most traits are, by definition, rare, the artificial virtues *cannot* be rare. For the conventions that constitute justice would not be stable unless most people were just (this does mean that *injustice* is by definition rare).[33] A second, related, point is that justice does not define someone as being of a special kind. It is not hard, after all, to be just (at least in Hume's property-oriented sense of 'just'), while generosity, wit, or the other natural virtues or vices require more of a person. This means that the indirect passions cannot be involved in identifying someone's justice. Third, since our concern with justice is not a concern with the kind of person someone is, the moral relevance of justice has to do, not with a person's general patterns of behavior, but with the particular acts he does. A generous person who takes someone else's property in order to aid those in need has still committed an injustice, despite his virtuous generosity.

Finally, I have noted throughout my discussion that Hume's analysis of character traits has taken the spectator's question 'How ought I to evaluate someone?' as its focus. But answering this question is easy in the case of justice, for it then asks only "Has this person obeyed the rules?" (T 3.2.6.7-8, SBN 529-531). The hard question about justice for a moral philosopher is why *ought* we to obey the rules (EPM 9.2.22-25, SBN 282-284). And this takes us back, not to the spectator's question, but to the deliberator's question 'What ought I to do'. The fourth difference between justice and the natural virtues, then, is that justice is a restriction that is supposed to be acknowledged by all of us as we decide what to do, while the natural virtues are tendencies to behave

[33] Hume even mentions justice when discussing vice and virtue as causes of the indirect passions (T 2.1.7.3, SBN 295) just pages after declaring that causes of the indirect passions must be rare (T 2.1.6.4, SBN 291). Hume might be meaning to distinguish between a minimal notion of justice as mere concurrence in the conventions and a more positive notion of justice as a virtue. Someone would be just in this stronger sense if she displayed an outstanding tendency to recognize and respect property rights, even, say, in the face of great temptation to cheat or steal. Hume does not hesitate to say that men "commonly" are "unjust and violent" (T 3.2.6.4, SBN 426; see T 3.2.9.3, SBN 551), by which he does not mean that the artifices are under constant attack, but rather that it is a rare occurrence for a person to be possessed of the virtue of justice in the stronger sense.

spontaneously in ways that have been taken up in our passional responses to one another.[34]

Acknowledgements

I would like to express my appreciation to Annette Baier, Stephen Engstrom, Sofia Reibetanz, Jacqueline Taylor, Sergio Tenenbaum, and Jennifer Whiting for their comments and criticisms on various versions of this paper. I presented parts of it at the Twenty-sixth International Hume Conference, Cork, Ireland, July 1999. Jane McIntyre was the commentator on that occasion, and provided me with very helpful and generous comments; I owe her thanks. I also presented versions of my argument at the Intermountain Seminar for Early Modern Philosophy, University of Colorado at Boulder, April 1998; and the Ontario Philosophical Society meeting in Kingston, Ont., Oct. 1998.

University of Toronto
Department of Philosophy
215 Huron St., 9th Floor
Toronto, Ont. M5S 1A2, Canada
e-mail: donald.ainslie@utoronto.ca

REFERENCES

Adams, R.M. (1985). Involuntary Sins. *Philosophical Review* **94**, 3-31.
Ainslie, D. (1995). The Problem of the National Self in Hume's Theory of Justice. *Hume Studies* **21**, 289-313.

[34] It might seem that my making the natural virtues depend on the emergence of a custom of feeling in society has erased the distinction between the artificial and natural virtues. For example, it is only insofar as there is a custom (or "maxim") of recognizing alms-giving as generous behavior that it is possible to act generously by so giving. And, indeed, this is a consequence of my view. But there is still this difference between the artificial and natural virtues as I understand them: There must be a co-ordination of sentiments in order for us to *recognize* a character trait; but there must be a co-ordination of actions for someone to *act* in line with the artificial virtues. Giving money to someone who is in need is an available act with or without our recognizing this act as typical of generosity; but taking someone else's property is not a *possible act* unless the artifices of justice are in place (hence the difference between the two kinds of intentions involved in the two kinds of virtue – see n. 11).

Ainslie, D. (1999). Scepticism About Persons in Book II of Hume's *Treatise*. *Journal of the History of Philosophy* **37**, 469-492.

Ainslie, D. (2005). Sympathy and the Unity of Hume's Idea of Self. In: J. Jenkins, J. Whiting and C. Williams (eds.), *Persons and Passions*, pp. 143-173. South Bend, IN: University of Notre Dame Press.

Alston, W.P. (1975). Traits, Consistency, and Conceptual Alternatives for Personality Theory. *Journal for the Theory of Social Behaviour* **5**, 17-48.

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

Aristotle (1985). *Nicomachean Ethics*. Indianapolis: Hackett.

Baier, A. (1991). *A Progress of Sentiments: Reflections on Hume's* Treatise. Cambridge, MA: Harvard University Press.

Brandt, R.B. (1970). Traits of Character: A Conceptual Analysis. *American Philosophical Quarterly* **7**, 23-37.

Bricke, J. (1974). Hume's Conception of Character. *The Southwest Journal of Philosophy* **5**, 107-113.

Butler, D. (1988). Character Traits in Explanation. *Philosophy and Phenomenological Research* **49**, 215-238.

Dees, R. (1997). Hume on the Characters of Virtue. *Journal of the History of Philosophy* **35**, 45-64.

Greig, J.Y.T., ed. (1932). *Letters of David Hume*, vol. 1. Oxford: Clarendon Press.

Hume, D. (1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Edited by P.H. Nidditch (3rd edition) and L.A. Selby-Bigge. Oxford: Clarendon Press.

Hume, D. (1978). *Treatise of Human Nature*. Edited by P.H. Nidditch (2nd edition) and L.A. Selby-Bigge. Oxford: Clarendon Press.

Hume, D. (1983). *History of England*, vol. 1 & 5. Indianapolis: LibertyClassics.

Hume, D. (1987). *Essays: Moral, Political, and Literary*. Revised edition. Edited by E.F. Miller . Indianapolis: Liberty Classics.

Hume, D. (1999). *An Enquiry Concerning the Principles of Morals: A Critical Edition*. Edited by Tom Beauchamp. Oxford: Clarendon Press.

Hume, D. (2000a). *Treatise of Human Nature*. Edited by David F. Norton and Mary J. Norton. Oxford: Oxford University Press.

Hume, D. (2000b). *An Enquiry concerning Human Understanding: A Critical Edition*. Edited by Tom Beauchamp. Oxford: Clarendon Press.

Kant, I. (1983). *Ethical Philosophy*. Translated by J. Ellington. Indianapolis: Hackett.

Kant, I. (1993). *Grounding for the Metaphysics of Morals*. 3rd edition. Translated by J.W. Ellington. Indianapolis: Hackett.

Korsgaard, C. (1996). Morality as Freedom. In: *Creating the Kingdom of Ends*, pp. 159-187. Cambridge: Cambridge University Press.

McIntyre, J. (1990). Character: A Humean Account. *History of Philosophy Quarterly* **7**, 193-206.

Norton, D. (1982). *David Hume: Common-Sense Moralist, Sceptical Metaphysician*. Princeton: Princeton University Press.

Pitson, T. (1996). Sympathy and Other Selves. *Hume Studies* **12**, 255-271.

Sabini, J. and M. Silver (1982). Character: The Moral Aspects of Traits. In: *Moralities of Everyday Life*, pp. 141-162. Oxford: Oxford University Press.

Sayre-McCord, G. (1994). On Why Hume's "General Point of View" Isn't Ideal – and Shouldn't Be. *Social Philosophy and Policy* **11**, 202-228.

Shaftesbury (1900). *Characteristics of Men, Manners, Opinions, Times, etc.*, vol. 2. Edited by J. M. Robertson. London: Grant Richards.

Wiggins, D. (1987). A Sensible Subjectivism? In: *Need, Values, Truth*, pp. 185-214. Oxford: Blackwell.
Williams, B. (1985). *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.
Wright, G. Von (1993). *The Varieties of Goodness*. Bristol: Thommes.

Stephen Engstrom

KANT ON THE AGREEABLE AND THE GOOD

One of the most notable features of Kant's practical philosophy is the sharp line it draws between the imperatives of practical reason, on the one hand, and the sensible inclinations, on the other, and the related way in which it contrasts their respective objects, the good and the agreeable. Though Kant allows that these rational and sensible ingredients are found mixed together in our ordinary practical life, he holds that the philosopher, like a chemist, can carry out an analysis that reveals them to be distinct elements. Indeed, though he emphasizes that there is no inherent opposition between them, the distinguished items are, in his depiction of them, so radically separate and heterogeneous from one another that they seem altogether unrelated. Virtue is no path to happiness, he maintains, nor is happiness a mark of virtue. However tangled and confused our rational and sensible motives may be in experience, however closely our inclinations, when suitably pruned and cultivated, may come to resemble requirements of reason, and however difficult it may consequently be to confirm, in any given case, the presence of a pure rational motive amidst the profuse abundance of sensible desire, these branches, Kant seems to be telling us, share no common root.

Understandably, this apparent separation has seemed questionable to many readers. If the rational and sensible elements are thus distinct and unconnected, it seems impossible to comprehend how they could ever be present together in a unified practical life. Kant places us at a crossroads: on the one side we have reason, admonishing us to follow the path of duty, goodness, and virtue; on the other we have the inclinations, beckoning us toward the agreeable and its pleasurable enjoyment. Faced with this stark display of options, readers have not surprisingly asked whether it is not possible to travel down both paths at once, and whether the map Kant has drawn can really be accurate if it displays these routes as leading to wholly different and unrelated destinations.

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 111-160. Amsterdam/New York, NY: Rodopi, 2007.

One familiar version of this general query is the much discussed question concerning the famously puzzling discussion of moral worth in the *Groundwork for the Metaphysics of Morals*, the question whether Kant can in any way accommodate inclination as a motive figuring in morally worthy action. If Kant insists that moral worth is found only in actions done from duty, never in those done from inclination, not even if they are in accordance with duty, will he allow that an action can be done from duty and so have moral worth if the agent also has an inclination to perform it? This question and others similar to it have prompted some fruitful investigations of Kant's view of the inclinations as well as kindred explorations of his view of the emotions and passions and his account of the role that sympathy and other feelings play in the conduct of morality and virtue.[1]

But a different aspect of the general concern comes into view when we observe that Kant conceives of practical reason as a cognitive faculty, a capacity to have a certain type of knowledge. This knowledge, which he calls practical knowledge, is the knowledge of what we ought to do, or of what it would be good, or right, to do, by which we direct ourselves in action; in its perfected form, it is the practical wisdom on which a virtuous character is based, but it is present in every moral subject, however much it may be obscured or impeded in the variable conditions of practical life. The common tendency among philosophers to conceive of knowledge as exclusively theoretical (or even as exclusively theoretical and empirical) can make it easy to suppose that when Kant speaks of practical reason he is conceiving of reason as employed for some other-than-cognitive purpose, a supposition that receives encouragement from the prominent role the idea of an imperative plays in his account of practical reason's principles, from the fundamental place the ideas of legislation and autonomy occupy in his ethics, and not least from his identification of practical reason itself with the will, a faculty whose exercise consists in the rational determination, not of belief, but of choice. Thus, we can easily come to think that Kant sees practical reason as concerned with action and its direction rather than with knowledge. But while there is no denying that he takes practical reason to be the source of laws and commands for action, it is equally true that he sees this legislative activity as cognitive in nature and as having its own distinctive object. Since the cognition is practical rather than theoretical, however, this object is known, not under the heading of the real, but

[1] To cite just a few recent contributions to the extensive literature on these questions: Herman (1993, Chs. 1, 2), Baron (1995, Chs. 5, 6), Korsgaard (1997), and Sherman (1997, Ch. 4).

under the heading of the good; or as Kant puts it, practical cognition is knowledge, not of "what is," but of "what ought to be." Once we recognize that practical reason is a cognitive faculty, it becomes clear that addressing the general question of how, if at all, practical reason's imperatives are related to the inclinations involves, in part, a consideration of the *objects* of inclination and reason. It becomes necessary to ask whether, and if so how and to what extent, or under what conditions, the object of inclination can itself be an object of practical reason and counted among the things that are good.

Given the way Kant separates these objects, and given that he occasionally seems to betray a disparaging attitude toward the inclinations, as in his often cited remark that "it must be the universal wish of every rational being to be completely free of them" (G 428),[2] it might appear that he thinks the agreeable can never be an object of practical reason. Such an appearance is bound to occasion alarm. Surely, cries out the hedonist in us all, there must be at least some connection, some real affinity, between the pleasant and the good! Kant points out that in common speech we readily mark the difference and allow that there are many things that are not good even though we greatly enjoy them. But in explaining such judgments he implicitly acknowledges that, in many cases at least, it is not in consideration of the enjoyment we derive from these things, but rather by looking to the effects, that reason declares them not to be good. Those who concede that they should forego their favorite dessert may do so because they think it to be too costly, or unhealthy, or addictive, or even because they think they deserve to be deprived of it on account of yesterday's excessive indulgence, but never simply because they find it delicious. Not even those who scorn the pleasures of the senses can despise them on such a basis. Moreover, Kant himself, a great admirer of Epicurus, does not hesitate to include happiness, which he conceives of as lying in enjoyable activities, as a necessary element in the highest good. There is some reason, therefore, to examine Kant's accounts of the agreeable and the good more closely to determine whether, on his view, the agreeable can ever really be deemed good, and if so how, given the way he separates them off from one

[2] References to Kant's writings are given by abbreviated title: Anth (*Anthropology from a Pragmatic Point of View*); G (*Groundwork for the Metaphysics of Morals*); KpV (*Critique of Practical Reason*); KrV (*Critique of Pure Reason*); KU (*Critique of the Power of Judgment*); MS (*Metaphysics of Morals*); R (*Religion within the Boundaries of Mere Reason*). References to the *Critique of Pure Reason* use the page numbers of the first (A) and second (B) editions; all other page references cite the page in the appropriate volume of Kant (1902-). Translations are my own, though I have consulted the commonly used English translations.

another. Such an investigation should yield a better understanding of how the agreeable and the good differ, and of how they are nevertheless related. It will also throw light on Kant's view of pleasure.

1. Rational and Sensible Desire

1. Let us start by comparing the representations that have the good and the agreeable as their objects. On the one side, there is the exercise of practical reason, which, in view of Kant's identification of the will with practical reason, we can alternatively characterize as the action of the will. On the other, there is inclination, or sensible desire. Before we attempt to determine what Kant's basis is for drawing this distinction, we should consider what, if anything, he thinks inclinations and acts of the will have in common. One indication that Kant supposes there to be an affinity between them is that he regards both as species of desire and characterizes the will itself as a faculty of desire, namely, "the faculty of desire whose inner determining ground [. . .] is met with in the subject's reason" (MS 213). Given this characterization of the will, we can describe the will's action as *rational desire*.[3] In addition, it is significant that Kant also speaks of a single "faculty of desire" to which these two species of desire – rational and sensible – belong. If both acts of the will and inclinations spring from a single capacity, then they must have something in common. We can locate this shared element by considering Kant's definition of this faculty. The faculty of desire, he states, is a living being's "capacity [*Vermögen*] to be through its representations the cause of the actuality of the objects of these representations" (KpV 9n).[4]

[3] Though the expression 'rational desire' does not correspond exactly to any standardly employed phrase in Kant's philosophical terminology, it will be useful to employ it here to facilitate comparison between the two species of desire. Other considerations supporting this usage include Kant's identification of the will with practical reason, his designation of the other species as "sensible desire" (*sinnliche Begierde*), and his awareness that his distinction is a continuation of the traditional scholastic distinction between *appetitus rationalis* and *appetitus sensitivus*. These considerations do not, of course, indicate that it is an arbitrary decision on Kant's part to speak of practical reason instead of rational desire. If, as he argues, reason by itself can move us to act, then reason itself has an efficacious power – a fact that is well captured by the phrase 'practical reason' but not so clearly expressed by the ambiguous 'rational desire', which might be read as meaning merely desire *in accordance with* reason, rather than desire *from* reason (cf. §1.2 below).

[4] This definition – indeed the very idea of a *faculty* of desire, which entails that desire is the exercise of an active power or capacity – may seem odd, given the hoary traditional association of desire with need, defect, and limitation (an association clearly expressed in

This definition suggests that desire can be characterized as a species of representation, and that what is distinctive about it is the causal relation it bears to the object it represents. This characterization can be elucidated by briefly considering how it applies to the two species of desire, rational and sensible.

Since rational desire lies in the exercise of practical reason, which is a cognitive power, and since the exercise of a cognitive power must lie in cognition, or a type of cognition, rational desire must be practical cognition, the cognition of practical reason. Thus, to understand the desiderative character of rational desire it will be necessary to consider the practicality of practical cognition, or what distinguishes the cognition of practical reason from that of theoretical reason.

Today the distinction between the theoretical and the practical is liable to be understood in terms of a difference in "direction of fit" between representation and object.[5] According to this way of thinking, in the theoretical case the representation is supposed to fit the object (or "the world"), while in the practical case the object is supposed to fit the representation. If some theoretical representation and its object do not agree, it is the representation, not the object, that is to be regarded as failing to fit and that must be adjusted or replaced in order to secure the requisite relation of fit. But where there is a lack of agreement between some practical representation and its object, it is the object, not the representation, that fails to fit and so needs adjustment or alteration. Thus, theoretical representation is thought of as standing to its object as an artist's sketch of a building stands to the building sketched, and practical representation is regarded as standing to its object as the architect's plan for a building stands to the building constructed according to it.

Kant's distinction between theoretical and practical cognition superficially resembles this distinction but is in fact quite different. To

---

such words as 'want' and 'passion'). For Kant, the original conception of desire is that of the exercise of a productive power, or a capacity to produce effects (a capacity that is itself an expression of lifepower: KpV 23), even if there is a species – sensible desire – that also implies need.

[5] Those who conceive of this distinction in terms of a difference in direction of fit also commonly conceive of it as a distinction between belief and desire, or between cognitive and non-cognitive (or conative) representation. See, for example, Dancy (1993, pp. 27*ff*), and Smith (1994, pp. 111*ff*). It is to be noted that these broadly Humean ways of conceiving of the distinction have built into them from the start a conception of knowledge as theoretical, and as a result they make it difficult to see practical knowledge even as a possibility. (At KpV 9n Kant objects to a way of defining the faculty of desire familiar in his own day that is tendentious in a corresponding way.)

accommodate the possibility of synthetic *a priori* knowledge, Kant holds that the objects of our cognition must conform to our cognition of them. He therefore has no place for the idea of cognition's fitting its object, so the difference between theoretical and practical knowledge, as Kant understands it, cannot lie in a difference in direction of fit. Instead, Kant draws the distinction in terms of a difference in the direction of existential dependence. Theoretical knowledge, as knowledge of "what is," is of independently existing objects that must be given by way of the senses in order to be known, whereas practical knowledge, or knowledge of "what ought to be," is knowledge that produces its objects, or makes them actual (KrV Bix-x; cf. KpV 46, 89). Hence, theoretical cognition depends for its actuality upon the actuality of its object, whereas in the case of practical cognition the actuality of the object depends on the actuality of the cognition. The theoretical knowledge we have of the stars, for example, depends on the stars' existence. Were there no stars, there would be no theoretical knowledge of them.[6] But on the other hand the astronomical observations and experiments we make in order to learn about the stars are brought about by our practical knowledge that it would be good to observe and to investigate them, and likewise the existence of the telescopes and other devices we use in these investigations depends on our practical knowledge of what is useful to our scientific purpose, in accordance with which we construct, maintain, and employ these instruments. Similarly, a virtuous person's actions of promise-keeping and helping others in need are the result of the practical knowledge such a person has that such things should be done. Without practical knowledge of them as things to be done or to be produced, none of these actions and objects would take place or exist. To the extent that they do, of course, they belong to the realm of the actual – "what is" – and can be known theoretically.[7] Thus, while theoretical and practical

[6] The dependence intended here is real, not merely conceptual or logical. The concept of knowledge guarantees that from its being known that stars exist it follows that stars exist, and hence that if stars did not exist, there could be no knowledge that they do. But there is also a real dependence of actual knowledge that stars exist upon actual stars: the latter must be, as Kant says, "given" to us in order for us to know of them, and they are given to us through their being present to the senses, whereby they somehow (directly or indirectly) affect the mind. This relation of real dependence holds, in a general and indirect way, even where the cognition comes first in time, as it does in the case of the man who, having undermined the foundations of his house, knows *a priori* (in a comparative sense) that it will collapse, for such knowledge depends on general knowledge, gained from prior experience, that bodies are heavy and therefore fall when their supports are removed (KrV B2).

[7] Such cognition, however, is not the *purely* theoretical cognition investigated in the *Critique of Pure Reason*, but rather theoretical cognition in which certain concepts

cognition differ essentially in the relation in which they stand to their objects, they may nevertheless share the same object between them: the practical knowledge that one ought to keep one's promises and the theoretical knowledge that one does keep them have a common object and differ only in respect of the difference in the direction of the existential dependence constituting their relation to it.

What distinguishes practical cognition from theoretical, then, is its efficacy, by which it works to produce its object, or to make it actual. Accordingly, our capacity for practical knowledge, as a capacity to be through our cognition the cause of the actuality of the object of this cognition, is a specific form of the general "capacity to be through [our] representations the cause of the actuality of the objects of these representations" and therefore belongs to the faculty of desire.[8]

Kant's distinction between theoretical and practical cognition is a distinction between two species of rational cognition and hence between two species of representation produced by reason. But a parallel distinction, or rather another instance of the same fundamental distinction, can be found among sensible representations. This distinction too is based in a difference in direction of existential dependence, but it applies to the sensible representations that can figure in, or contribute to, cognition. In the theoretical case, these are sense perceptions, or, more precisely, sensations, which according to Kant constitute the matter of sense perceptions (as opposed to the form, which lies in intuition). In the practical case, they are sensible desires. As the effect produced by an object on the mind through one of the outer senses, a sensation depends for its actuality on the actuality of the object. A sensible desire, in contrast, has an efficacy through which it works to produce its object, whereby the actuality of the object depends on the actuality of the representation.

Thus the sensation of red of which I am aware when I look at an apple depends on the presence of the apple, as an object affecting the sense of sight. Were the apple not present to my senses, I would not have this

originating in practical cognition also figure – notably the idea of the free practical subject that constitutes both the form of practical self-consciousness and the subject of practical knowledge.

[8] In Kant's usage the term 'practical' normally serves to mark the distinctive efficacy of reason; it thus expresses the efficacy of rational desire (efficacious cognition), rather than that of desire in general (efficacious representation). Kant does, however, occasionally use this expression in an extended sense, in relation to the faculty of desire in general (as when he speaks of "practical pleasure": see §4.2 below). The difference between practicality, or the efficacy of rational desire, and the efficacy of sensible desire will be briefly considered below (§1.3).

sensation of red. But if I am moved by a sensible desire to eat the apple, then it is the presence of the object – that is, the presence of the apple to my senses of taste, smell, and touch through my action of eating it – that depends on the representation that constitutes the desire (i.e., the representation of its agreeable qualities – its taste, smell, and so forth).[9] Were it not for the desire, I would not be eating the apple, nor therefore would the apple be present to these senses.

It is important to note that the basis for classifying a desire as a *sensible* desire lies not in the representation itself (i.e., not in anything pertaining merely to its representational function), but rather in the origin and character of its efficacy. If it were the character of the representation taken by itself rather than that of its efficacy that determined whether a desiderative representation is sensible, only representations of sensible qualities (such as those involved in the enjoyment of an apple) could qualify. But while the representations of pleasing sensible qualities provide the primitive case of sensible desire, Kant emphasizes that other representations can also count, including efficacious thoughts arising from theoretical cognition.[10] Persons who were first moved to help others from sympathy, for example, may, independently of their practical rational cognition that they ought to help others, find pleasing the awareness they come to have of their own helpful action as they witness its successful outcome, and accordingly this theoretical knowledge of their helpfulness on such occasions may itself give rise to an inclination to be helpful. How sensible desire arises in these cases will be considered more closely below (§1.2).

In sum, what rational and sensible desire share in common in virtue of their belonging to a common faculty of desire lies, not in the contents, or the objects, of the representations in which they consist, nor in a direction of fit, but rather in their efficacy. Desires of both types count as desires, as representations belonging to the faculty of desire, on account of their causality in respect of their objects, a causality through which those objects will be produced, or brought into existence, provided of course that the conditions are right. (Obviously desire may fail to realize its object if there is something opposed to it that blocks its efficacy, or if

[9] Here and in what follows the context should indicate in each case whether 'taste', 'smell', etc. are being used to refer to the specific mode of the power to represent (sense), its operation (sensation), or the object (sensible quality).

[10] See KpV 23, MS 212n. The error of supposing the difference between rational and sensible desire to lie in the representations (rather than their efficacy) receives extended criticism in Kant's well-known discussion of how the higher and the lower faculties of desire are to be distinguished (KpV 22-25).

the conditions on which its efficacy depends are undermined or otherwise absent, just as a plant may fail to grow if afflicted by disease or deprived of water and light.)

2. Kant traces the difference between rational and sensible desire to the difference in their origin, to the difference between reason and sensibility. This latter distinction is, of course, a general one that runs through both Kant's theoretical and his practical philosophy. He locates its basis in the difference between spontaneity, or the cognizing subject's capacity to act from itself, and receptivity, its capacity to come to have certain affections, that is, its capacity to take on, or to "receive," certain states, or determinations of the mind, through being affected by things existing outside itself.[11] Thus, in the case of theoretical cognition, Kant bases the distinction between the rational and the sensible *representations* that figure in such cognition in the difference between the subject's capacity to produce from itself representations of objects, and its capacity to come to have (or to "receive"), through being affected by certain objects existing outside it, representations of such objects.[12] Now as we have noted, the faculty of desire as Kant defines it is the capacity to be through one's representations the cause of the actuality of the objects of those representations. The difference between rational and sensible *desire* will accordingly be based in the difference between the spontaneous and the receptive capacity to be such a cause – the difference, that is, between the capacity *to be from oneself* through one's representations the cause of their objects' actuality, and the capacity *to come, through being affected by certain objects, to be* through one's representations of such objects the cause of their actuality. Let us consider each of these types of desire more closely.

Implicit in the characterization of the capacity of rational desire just provided is the idea that if a cognizing subject is from itself, or spontaneously, through certain of its representations the cause of their objects' actuality, then it must also be the source of the representations through which it is this cause. It seems clear that this must be so. If the

[11] To avoid unnecessary complications, it will be convenient here to employ this restricted conception of receptivity in lieu of Kant's broader conception, which does not specify that the affection is by things existing outside the cognizing subject and so leaves room for a type of receptivity involving self-affection.

[12] KrV A51/B75. Though theoretical cognition is knowledge of objects given to the cognizing subject from elsewhere, and hence cognition whose actuality depends on the actuality of its object, Kant holds that the representations in which such cognition consists are nevertheless spontaneous in virtue of being based in certain intellectual conditions of the possibility of all such cognition of those objects, as appearances.

subject's having these representations depended on its having been affected by objects existing outside it, then so would the subject's capacity to be a cause of the actuality of the objects of these representations. This dependence on affection would thus preclude the absolute spontaneity that belongs to the type of desire that has the connection with reason implied by the idea of rational desire, as desire *from* reason.[13] Only through spontaneity in its representation can a desire be traced to rational cognition. The following observations, though brief, may help bring this connection into clearer view.

For a cognizing subject to be conscious of the diverse representations figuring in its cognition as representations it produces from itself is for it to be conscious of them as specifications of fundamental representations that are necessarily ingredient in all possible cognition and therefore not dependent upon the subject's having been affected by external objects. As the starting point for the act of cognition, these fundamental representations are principles of cognition, to which everything else belonging to cognition must conform. Such principles, and the cognition derived from them, are exactly what Kant identifies rational cognition as consisting in. Knowledge from reason, he holds, is just knowledge from principles.[14] Insofar as desire is rational, therefore, it has its source in cognition from principles and is nothing other than such cognition so far as this cognition is practical, or productive in respect of its object. This practical cognition, as was noted earlier, is the knowledge of what we

[13] At first glance, it might appear possible for there to be a *relative* practical spontaneity, which would belong to a subject that had the capacity to produce from itself the actuality of the objects of representations received from elsewhere. Such a capacity, if possible, would be the source of a third species of desire, a hybrid distinct from and intermediate between rational and sensible. But this apparent possibility dissolves in the face of the consideration that since all spontaneity depends on self-consciousness, which in turn is found only in the activity of conscious representation, all spontaneity is originally spontaneity of representation. The capacity to produce from oneself the actuality of the object of a representation thus depends on the capacity to produce from oneself the representation as well. To rule out this hybrid possibility is not, however, to rule out the possibility – to be described more fully below (§5) – that certain self-produced representations belonging to rational desire may themselves depend on material conditions furnished by representations received from objects existing outside the subject. Nor is it to rule out the possibility of occasions on which "reason stands in merely as a handmaid of the natural inclinations" (R 45n): the representation of action that constitutes an intention to pursue an object of sensible desire is always an exercise of thought or reason, even where the efficacy of that representation has its source in sensible rather than rational desire.

[14] KrV A299-300/B356-57. Cf. KpV 12: "we say that we cognize something through reason only when we are conscious that we also could have known it even if we had not thus encountered it in experience."

ought to do, and its principle and form, Kant holds, is reason's *a priori* idea of practical law.[15] (Why this cognition is necessarily a cognition of *action* – what we are to do – will emerge below, when we consider the distinctive character of its efficacy [§1.3].) The capacity of rational desire, then, is the capacity to act from principles of rational cognition, or from the conception of law, and this capacity Kant identifies with the will, or practical reason (G 412). Thus, in rational desire the subject is spontaneous both in respect of the representation (rational cognition) and in respect of the latter's efficacy: reason, as a cognitive faculty, spontaneously gives rise to cognition in accordance with its principles of rational cognition in general (first moment of spontaneity), and where it spontaneously determines this cognition as practical (second moment of spontaneity) its act is an act of will, or of rational desire, which can (and *will*, if nothing interferes) determine the free power of choice to choose actions through which the object of practical cognition can be realized.

In a corresponding way, sensible desire is doubly receptive. As we noted, sensible desire is based in a subject's capacity to come, through being affected by certain objects, to be through its representations of such objects the cause of their actuality. The affection mentioned here always has two sides, one pertaining to sensible desire's representation of its object, the other pertaining to this representation's efficacy. Let us consider again the primitive case of a sensible desire for an apple. On the one side, there is receptivity in respect of the *representation*, since the representation is acquired through the subject's being affected by objects existing outside it. A sensible desire for an apple depends on one's having already tasted one: the representations of the sensible qualities (the tastes and smells) of an apple that figure in this sensible desire are not produced from oneself, but depend on antecedent sensations produced in the mind through the presence of an apple to the senses. But on the other side, the subject is also receptive in respect of its being the *cause* of the actuality of the object of this representation. For the subject is the cause only insofar as the faculty of desire (the subject's capacity to be such a cause) comes to be determined by antecedent representations (in this case sensations) that are had through the subject's being affected by their objects, this determination of the faculty of desire being something of which the subject is conscious through its capacity to feel pleasure and pain. It is because an apple's tastes and smells are pleasing (because, in other words, the sensations produced by the apple affect the faculty of

[15] Kant presents his account of this form of cognition – which he characterizes as the "form of a universal legislation" – and the role it plays in grounding the practical laws in which practical cognition consists in the *Critique of Practical Reason* (KpV, §§1-7).

desire in the positive manner that is inwardly manifested in consciousness as the feeling of pleasure) that one "receives," or comes to have, the sensible desire for an apple. Thus, sensible desire arises through the felt affection of the faculty of desire by representations that in turn are had only through the mind's being affected by the objects they represent.

It is important, however, not to overlook the point mentioned earlier, that sensible desire need not arise from sensations, at least not in the immediate way described in the example just given. A suitably general characterization of the origin of such desire can be achieved by using, in place of the notion of a sensation, the more general notion of a *representation of the existence (or actuality) of a thing*, which Kant sometimes employs in this connection (e.g., at KpV 22; cf. KU 204). Thus, we can say that sensible desire springs from the pleasure found in the representation of the existence of some object. It is clear from what has just been said that this more general terminology is not to be understood in a sense so broad as to embrace the mere fanciful imagining of a thing as existing or actual ("Wouldn't it be wonderful if . . . ?"). Such wishful imagining is not the original source of sensible desire, but rather one way in which such desire may manifest its presence. As it will be used here, the expression 'representation of the existence of a thing' signifies a representation that stands to its object in the relation of existential dependence that distinguishes theoretical cognition, as knowledge of "what is," from practical cognition, regardless of whether the representation by itself amounts to theoretical cognition. There are two types of representation of the existence of a thing that should be distinguished:

(i) sensations, or representations immediately produced in the mind by existing things (including the kinesthetic sensations arising from movements of one's body as well as other outer sensations, such as the tastes and smells produced by an apple),

(ii) theoretical cognitions that concern actual things and determine how they actually are, judgments that are always empirical and therefore themselves depend, directly or indirectly, on the first sort of representation of existence (sensation).

Though sensations constitute the primitive source of sensible desire and perhaps in part for this reason receive a large share of attention when Kant discusses desires of this type, he also recognizes the very considerable extent to which sensible desires arise from the theoretical cognizance of one's actions and achievements (cf. KpV 23-24). One

example of a desire that springs from such cognition is the immediate inclination to help others that Kant attributes to the sympathetic philanthropist he discusses in the *Groundwork* (G 398). In saying that this man finds an "inner gratification" in spreading joy around him, Kant implies that this inclination arises, not from any pleasing sensation, but from his immediately gratifying recognition of success in his efforts to help others. In fact, pleasure that stems from the awareness of the successful exercise of one's powers in the pursuit of this or that object may come to constitute a very great part of the basis on which a person's sensible desires are founded.[16] In the primitive cases the decision to pursue an object depends on the antecedent enjoyment of such an object's sensible qualities. (We may assume, for instance, that sympathetic pleasure and pain arising respectively from the sensations figuring in the experience of others' joy and sorrow first led to the helping actions that the philanthropist subsequently found immediately gratifying [or, to express the thought in Kant's terms, that it was *Mitleid* that first led to the philanthropist's *Teilnehmung*].) But the decision to pursue the object is the formation of an intention, which, though its efficacy depends on the presence of the sensible desire for the object, includes a new ingredient not originally contained in that desire, namely, a representation of the action or the pursuit; if one is successful in the pursuit, one's recognition of success is a new source of pleasure, and from the resulting pleasure a new desire, distinct from the original, is spawned, a desire that has such pursuit as its object.

Moreover, pleasure attending such recognition can also, to the extent that one becomes at least implicitly conscious that the awareness of success in the actions one undertakes is itself *in general* a source of pleasure, give rise to a general interest in the cultivation of one's powers and thereby to the adoption of objects of pursuit the representation of whose bare actuality may involve no pleasing sensations at all – aims and goals adopted solely on account of the pleasure expected to stem from the consciousness of their successful attainment. In this way, it becomes possible for the choice of the object to be determined less by the expectation that the object itself will be pleasing to the senses than by an estimate of the difficulty of its attainment (it must be difficult, but not too difficult). Indeed, the object need not be one the representation of whose existence involves any pleasing sensation at all; it may be nothing but the performance itself, as in displays of ingenuity or skill, and moreover one undertaken at least in part because the sensations and

[16] The basis of this pleasure will be considered below in §4.3.

thoughts it will involve are painful or unpleasant, as in feats of great strength, endurance, or fortitude.

The theoretical judgments from which sensible desires can arise need not concern objects of the outer senses; they may be judgments about oneself and one's own mental activities. All representations, even sensations, are mental activities, and insofar as these involve spontaneity one's engagement in them may be accompanied by a feeling of ease and facility or hindrance and difficulty – that is to say, a feeling of pleasure or displeasure attending the empirical consciousness of success or failure in the exercise of the powers of representation. Sensible desires can thus arise that have as their objects the types of mental activity (the exercise of wit, for example) in which one has enjoyed such consciousness of success. (The heightened or facilitated exercise of the powers of representation may of course itself depend in various ways on external objects and conditions, such as the presence of the necessary materials, or the interest or approval of other persons. This dependence on outer objects need not be mediated through the outer senses or theoretical judgments at all, as when the exercise of the imagination or thought is stimulated through the efficacy of an intoxicant, such as alcohol [cf. Anth §29], or by other artificial means, though if, as is commonly the case, one is aware of this dependence, a sensible desire for that object will result.)

In every case, however, the source of the efficacy of the representation in which sensible desire consists (the efficacy in virtue of which the representation counts as desire at all) is not a spontaneous act of cognition, but a determination of the faculty of desire by a representation that, being a representation of an object's existence and hence dependent on its object's actuality rather than productive of it, does not itself belong to the faculty of desire – a determination inwardly manifested in *feeling*, the sensible (receptive) awareness of the state (or activity) of mind[17] to which the representation of the object's existence belongs. This feeling, which Kant calls gratification (*Vergnügen*), is a species of pleasure found in the representation of the existence of some object.[18] Such an object – that is, one the representation of whose

[17] The expression 'state of mind' translates Kant's *Gemütszustand*. Since the *Gemüt* is "the principle of life" (KU 278), every such state is an activity comprising component activities, which may be in agreement or conflict.

[18] Kant uses more than one term to refer to this feeling. Besides *Vergnügen* (which might also be translated as 'satisfaction'), *Genuß* and its equivalent *Genießen* (enjoyment) are particularly worth mentioning. These terms are very similar in meaning, but *Vergnügen* carries a stronger suggestion of passivity, directing our attention back to the object that gratifies ("makes itself pleasing to") the subject through giving rise to a representation of its existence that affects the faculty of desire, while *Genuß* and *Genießen* (which have an

existence is accompanied by this feeling of gratification – is what Kant normally intends when he speaks of "the agreeable." On account of this feeling involved in the representation of its existence, the agreeable can be described as the "spring" or "motive" (*Triebfeder*) of sensible desire (KU 266).

Why this determination of the faculty of desire by a representation of the existence of a thing should be manifested in a feeling of pleasure is a question we are not yet in a position to address. But we shall return to it when we examine this feeling more closely below (§4) and take note of what, according to Kant, the awareness in which it consists is an awareness of.

3. The difference in origin just described is the basis of a formal difference between rational and sensible desire, a difference in the mode or type of efficacy distinctive of each. This difference can be succinctly expressed by saying that rational desire is productively related to its object only through its consciousness of this relation, whereas in sensible desire the awareness of the relation depends on the relation itself.

That rational desire's efficacy depends on its own consciousness of its efficacy can be seen by noting that the efficacy of rational desire is the efficacy of practical knowledge (the efficacy distinguishing it from theoretical), and that the efficacy of practical knowledge necessarily falls within the scope of such cognition's own self-consciousness. If the efficacy of practical knowledge did not fall within the scope of self-consciousness, then it would be possible to be engaged in practical judgment without any awareness, however obscure or implicit, that one was so engaged. But this is impossible. The distinction between practical and theoretical knowledge must be internal, in the sense that it is constituted spontaneously, through cognition's self-conscious self-determination. Whether an act of cognition is theoretical or practical cannot be determined by anything external to the awareness that constitutes the cognition itself, as if it could ever be the case that the

---

etymological connection with *Nutzen*, 'use') suggest that the object is already an object of the faculty of desire, that the subject is already engaged in the pursuit and use of it. First we find the object to be gratifying, then we pursue, use, and enjoy it. The opposite of gratification is pain (*Schmerz*). (Like Plato, Kant sees gratification and pain as so intimately connected – "two beings attached to one head" [Plato *Phaedo*, 60b] – that any reasonably comprehensive discussion of the former would require examination of its relation to the latter; indeed, he sees the latter as, in a certain respect, primitive: "pain must precede every gratification" [Anth 231]. For the limited purposes of this essay, however, attention will be directed almost exclusively to the positive member, as our focus throughout is on desire, the good, and the agreeable rather than their opposites, aversion, the bad, and the disagreeable.)

subject of practical cognition would require, in order to be aware that it was judging practically rather than theoretically, to undertake an empirical investigation, or to advert to anything outside its own activity. It is not by such means, nor indeed by any *means* at all, that in judging that one is (or ought) to do something one is aware that one is judging that one is (or ought) to do it, rather than judging (with reliance on sensation) that one will do it, is doing it, or has done it. Thus, practical cognition is necessarily aware of itself as practical. Moreover, it is practical at all only *through* its awareness of itself as such, for it is precisely through the cognitive act's self-conscious determination of itself as practical that it constitutes itself as practical. And since this practicality is just its efficacy, its efficacy is constituted by its awareness of itself as efficacious. This practical self-awareness is essential to all choice and all exercise of the will. The efficacy of choice, and of all rational desire, lies in its awareness, or understanding, of itself as efficacious. And because the awareness of this efficacy is the original practical representation of the action by which the represented object is to be produced, rational desire is essentially practical cognition of the *production* of the object, or cognition of action: cognition of "what ought to be" is always also cognition of what we ought to do.

To say that rational desire is efficacious only through its spontaneous understanding of itself as efficacious is not to say that to have a will is to have the power to produce anything and everything one might conceivably wish to bring about or fancifully conceive of oneself as bringing about. One's capacity to have a practical understanding of oneself as the cause of some represented object depends on certain conditions, notably one's own empirical theoretical judgments about the extent of one's capacities to produce effects. Cognizance of actual failure in one's previous attempts to achieve one's object erodes the confidence on which those first attempts depended, the confidence that the attainment of the object was possible. The ensuing doubts and scruples about one's abilities make it difficult to sustain the practical understanding of oneself as the cause and to that extent difficult to choose to try again. Similarly, a theoretical conviction that success is impossible undercuts the possibility of choice altogether (KpV 57). Kant sees this dependence of rational desire on theoretical representation as extending even to the case of pure practical reason's fundamental law, notwithstanding its independence from all empirical sources; his doctrine of the postulates of pure practical reason articulates the theoretical convictions on which, he argues, the practicality of this law depends.

In the case of sensible desire, on the other hand, the efficacy does not have its origin in a spontaneous act of cognition. Rather than being constituted by self-awareness, the efficacy of sensible desire is independent of, and hence prior to, the subject's awareness of it. Here the subject does not bring about the efficacy through an act of practical judgment, but instead is aware, through feeling, of an antecedently present efficacy. This sensible, or receptive, species of awareness of the efficacy of desire can vary widely from case to case in intensity, constancy, and pattern of recurrence, and it can arise in various ways: not only directly, in the pleasure felt when the efficacy is completed in the actualization of the object or in the displeasure felt when that pleasing actualization is interrupted or broken off, but also in the hopeful or fearful expectations regarding the prospect of the object's actualization, in the enlivening feelings of anticipation attending progress in its pursuit or the frustrations and disappointments of failure, and in the felt influence the desire has on one's thoughts and attention, as in the wishful imagining mentioned earlier. In all such cases the efficacy of the representation is prior to the awareness of it. It is through our capacity to feel pleasure and pain that we are aware of our inclinations, but feeling, according to its very idea, is not the mode of awareness that belongs to the self-determination of spontaneity and cognition.

It might be objected here that in fact the pleasure we find in the representations of the existence of an object through which the faculty of desire is first determined to a sensible desire for such an object provides us with an awareness of the sensible desire – and hence of its efficacy – that is prior to that efficacy. Does not the pleasing taste of the first bite into an apple precede the sensible desire for apples that arises through that experience? Kant himself, when speaking of the relation between pleasure and sensible desire, states that the faculty of desire is determined through pleasure (KpV 21) and that the pleasure precedes the determination as its cause (MS 212).

But while it is true that it is through such a pleasing representation that sensible desire first arises, the *feeling* of pleasure accompanying the representation is not itself a spontaneous act, but rather an awareness of the determination of the faculty of desire through which that sensible desire arises – an awareness that depends on that determination. And since this determination is itself the coming to be of the sensible desire, which, like all receptively acquired capacities, comes to be only through its exercise, it follows that insofar as that determination is prior to the feeling of pleasure that constitutes the awareness of it, so is the first exercise of the efficacy of the sensible desire that arises through that

determination. (As for Kant's statement that the pleasure precedes the determination as its cause, this need not be read as incompatible with the identification of the felt determination of the faculty of desire with that faculty's being determined to a sensible desire. Being determined depends on an act of determining, which is prior to it, even though the determining and the being determined are the same.)

4. To summarize what has come to light so far: Acts of the will and sensible desires have a generic affinity in that they are determinations of the general faculty of desire and as such are efficacious representations, or representations through which the subject is a cause of the actuality of their objects. But because Kant locates the distinction between rational and sensible desire in a difference in their sources – a spontaneous act of cognition in the one case, and in the other a representation that, though it may itself be a cognition, determines the faculty of desire in such a way that the subject can be aware of the determination only through the feeling of pleasure, which, like all feeling, belongs to receptivity – he thinks of these two sorts of desire as "heterogeneous," or "specifically different." Thus, rejecting the view of his Leibnizian contemporaries that the difference lies simply in a difference in the degree of clarity and distinctness with which the desired objects are represented,[19] Kant maintains instead that it is a difference in kind, reflecting a division of the faculty of desire itself into the "higher" and "lower" faculties of desire (KpV §3). No desire can be both rational and sensible, nor can a desire of one of these types develop into one of the other.

2. The Good and the Agreeable

1. Let us turn now to the objects of rational and sensible desire. It was stated earlier that these are, respectively, the good and the agreeable. We should first confirm that this is so by briefly considering what Kant says about our prephilosophical understanding of these objects. We will then be in a position to look for further differences between them that flow from the differences we have already noted between the two sorts of desire.

Kant states that our usage of the terms 'good' (*das Gute*) and 'agreeable' (*das Angenehme*) distinguishes the agreeable from the good

[19] See Beck (1960, pp. 94-95). Kant's division is of course opposed to *any* view of the difference that would see it as a matter of degree – e.g., Hume's depiction of it as between calm and violent passions (1739, Book 2, Part 3, Sects. 3, 8).

and demands that the good be "judged through reason, and therefore through concepts, which can be universally communicated, and not through mere sensation [or feeling],[20] which is restricted to individual subjects and their receptivity" (KpV 58; cf. KU §7). When we deem something to be good, such as a proposed action or the end it promotes, our judgment includes the implicit thought that there should be general agreement in the matter, that any rational subject is in principle capable of arriving at the same judgment (the same combination of the same concepts), and hence that all such subjects would in fact agree to the extent that they were informed and properly exercising their reason. When we find something to be agreeable, on the other hand, our judgment involves no such implicit thought; we are ready to allow that what is agreeable in our estimation may not be so in that of others, and may even be disagreeable to them. If two subjects differ over whether a given object is good, there is a genuine disagreement between them, whereas no true disagreement can arise where they speak only of what is agreeable or disagreeable. In the latter case the affirmation and denial are relative to the receptivity of the individual subjects; in the former they are not.

In pointing to this difference in form between judgments made through reason and hence through universally communicable concepts and judgments made through mere individual feeling, Kant is explicitly appealing to the distinction between spontaneity and receptivity discussed above, by which rational cognition and sensible representation are distinguished. By briefly revisiting these distinctions we can confirm that the good and the agreeable, as understood through this formal difference in the judgments regarding them, are, respectively, the objects of rational and sensible desire.

Rational cognition, we noted, is cognition from principles. But since rational cognition, according to its very idea, can be shared by all rational subjects, all such subjects must at least implicitly recognize, through the self-consciousness that such cognition involves, that this cognition can be universally shared and hence that they all share the same cognitive capacity and therefore also the same fundamental spontaneous representations that figure constitutively in their exercise of it and serve as the principles on which they rely when they judge through reason. The formal feature of universal communicability that Kant finds in judgments

[20] The term here is *Empfindung*, which Kant commonly uses to refer to sensations, or impressions of the outer senses. He points out, however, that this expression is ambiguous, and in some contexts he employs it to refer to the feeling of pleasure (or displeasure) (see KU §3).

of the good – judgments in which, as our usage of 'good' demands, the good is "judged through reason" – is accordingly a reflection of these judgments' implicitly self-conscious reliance on shared principles of reason and hence a mark of their status as cognitive judgments of reason. The good, as an object "judged through reason," is thus an object of rational knowledge.

In identifying the good as an object of knowledge, Kant is not, of course, denying that it is an object of desire. "What we are to call good," he claims, "must be in the judgment of every rational human being an object of the faculty of desire" (KpV 60-61); and he says we understand the good to be "a necessary object of the faculty of desire [. . .] according to a principle of reason" (KpV 58). If the good is such an object, it must be the object of the will – "a faculty of desire determined through reason" (KU 209) – or, equivalently, the object of practical reason and practical knowledge. The concept of the good, then, refers to what is represented in practical cognition, and this is the object of rational desire.

The agreeable, on the other hand, is judged through mere individual feeling. According to our preceding discussion, the object of sensible desire is the object of a "received" representation through which the subject is a cause of the actuality of that object. We have seen that this representation is received in that the subject comes to have it only insofar as the subject's faculty of desire is determined by an antecedent representation of the existence of a thing (a representation that depends on the subject's being affected by the represented object); we have also noted that the determination of the faculty of desire through which this representation (the sensible desire) is received is something of which the subject is conscious through a feeling of pleasure, or gratification, that it finds in the representation of the existence of that object. But since this pleasure by which the subject is aware of the determination of its faculty of desire is also at the same time a judgment "through mere feeling, which is restricted to individual subjects and their receptivity," a judgment in which that object is found agreeable, and since it is precisely such an object that is the object of the sensible desire that arises through the determination of the faculty of desire manifested in that feeling of pleasure, sensible desire has the agreeable as its object. (See, e.g., G 413, KpV 21-25, KU 205-210, 266.)

The good and the agreeable differ in form, then, in that they are picked out by terms that signify, respectively, the object of practical cognition (what is affirmed by reason) and the object of gratification (what is affirmed by the senses) and hence are objects of types of desire that themselves differ essentially, or in form, in virtue of the different

capacities of the subject – spontaneity and receptivity – from which they arise.

2. A further difference between these objects flows from the essential difference we found earlier (§1.3) between the types of efficacy characteristic of rational and sensible desire (a difference that, as we saw, stems in turn from the different origins of these two sorts of desire). This difference pertains to the essential relation the object bears to the desire for it: the object of rational desire is essentially the *effect*, whereas the object of sensible desire is essentially the *cause* or occasioning condition, of the representation in which the desire consists.

Because rational desire is efficacious only though being conscious of itself as efficacious, it always represents its object as its own effect. Thus the object of rational desire is essentially conceived by the practically cognizing subject as the effect it is to produce through some action, so that – as Kant famously points out (G 417) – to will something as one's end is to will it as one's effect and thereby to will the action required to produce it. Because the conception of the object is essentially the conception of an effect of the practically cognizing subject, the object cognized (which must be in agreement with the cognition of it) is likewise essentially an effect of the practically cognizing subject. The good, accordingly, is essentially the effect of practical knowledge. Nothing but practical knowledge can produce the good.

Since sensible desire does not involve this self-conscious efficacy, we cannot similarly conclude that the object of sensible desire is essentially the effect of sensible desire. On the contrary, the object of such desire is always something that, as Kant observes (G 401; cf. G 395, KU 431), could be produced by the hand of nature, rather than through our free choice and action. Indeed, were the objects of sensible desire not producible by other causes (instinct, for example), sensible desire itself could never arise. For if we consider again the account we have already outlined of the origin of sensible desire, it is clear that the relation such desire bears to its object is the reverse of the relation we have just seen to hold in the case of rational desire. The source of the sensible desire is not a spontaneous act of knowledge, but a determination of the faculty of desire manifested in the feeling of gratification, a species of pleasure found in the representation of the existence of the agreeable. The agreeable, in fact, is nothing but an object the representation of whose existence involves this feeling of gratification. Through this feeling, then, the agreeable is, as we noted earlier, the "spring" from which sensible desire arises (KU 266). Thus the object of sensible desire is necessarily

the cause, or at least the occasioning condition, from which that desire arises.

3. Because this difference between the causal relations the good and the agreeable bear respectively to rational and sensible desire is essential, or formal, in nature, it can also be expressed as a difference in the forms of these objects. Although the objects of rational desire may vary, they all share a common form in virtue of their relation of essential dependence on rational desire as their cause. We can accordingly distinguish between the form of these objects, which lies in this relation of causal dependence, from the matter or content, by which one such object can be distinguished from another. Since this formal relation of dependence is nothing other than the action by which the objects of rational desire are to be produced, and since such action, springing as it does from the spontaneity of rational cognition through the latter's determination of the subject's free power of choice, is always free, free action is an essential element of the good, indeed its very form.

Similarly, we can distinguish between the form and the matter of the object of sensible desire. Although the objects of such desire vary according to the variety in the antecedent representations of the existence of the things on which, through those representations, sensible desires depend, these objects all share a common form in virtue of the relation of essential causal dependence sensible desire has on its object. And since this dependence lies in the determination of the faculty of desire manifested in the pleasure – i.e., in the gratification – first felt in the antecedent representation of that object's existence, agreeableness, or the capacity of an object to produce (in the subject) gratifying representations of its existence, belongs to the essential form of the object of sensible desire, however much variety there may be in the matter or content.

So where we are concerned with the forms of these two species of object, rather than the differences in the matter by which instances belonging to one or the other of these species might be distinguished from one another, we can speak of the object of rational desire as *the effect of (free) action*, and that of sensible desire as *the cause of gratification*. Thus, as Kant observes, "good and bad are properly related to actions, not to the person's state of sensation [or feeling]" (KpV 60); it is the agreeable and the disagreeable that are properly related to the person's state of feeling: "the agreeable [. . .] as such represents the object solely in relation to the senses" (KU 208).

4. Despite the considerations just advanced in support of it, the foregoing characterization of the difference in form between the good and the agreeable will perhaps seem puzzling or problematic in certain respects. In particular, it might seem that the identification of the object of sensible desire with the cause of sensible desire conflicts with the account of desire in general outlined earlier, which proposed that for a subject to desire something is for it to be through one of its representations the cause of the actuality of the object of that representation. That characterization might appear to suggest that with regard to all desire, sensible desire included, we ought to be thinking of the object as the effect, not the cause.

But on closer inspection it is not difficult to see that there is no conflict here. All desire, whether rational or sensible, stands in an efficacious relation to its object, through which it produces the object, where conditions allow. But only in the case of rational desire is the object essentially one that can be produced *only* by the desire, and only in the case of sensible desire is the desire essentially one that can arise *only* through the effect its object produces on the subject. Thus, another way in which sensible desire's distinctive causal dependence on its object can be expressed is by saying that whereas rational desire is essentially *pro*ductive, sensible desire is essentially *re*productive.

This response assumes, of course, that some account of this reproductive character is available. But to attribute such a character to sensible desire will perhaps at first glance seem paradoxical, if not incomprehensible. How can sensible desire – or anything at all, for that matter – have its own cause as its effect?

It was not, however, a certain particular thing, or individual existing object, that was identified as the cause of sensible desire, but rather some object the representation of whose existence produces a feeling of gratification. The object was from the start understood to be the object of a representation belonging to a pleasing and hence self-sustaining state of mind (cf. §3.3 below) and as such the object of a representation that has a certain generality, since in order to belong to such a state of mind the representation of the object's existence must be ongoing and hence indefinitely repeatable: the sensible desire I now have for an apple, which arose through the pleasure attending the representations figuring in my eatings of apples in the past, does not have for its object any of those devoured and no longer existing apples, but rather an apple. It was not those particular apples that were needed to produce the sensible desire for an apple; other apples I might have eaten instead would have served as well. What was needed to produce the desire was just the eating of

some apple – apple eating, that is – in which the sensations involved were found pleasing. The cause of this sensible desire, then, was neither this apple as opposed to that one, nor that one as opposed to this, neither one apple nor many, but "some apple or other" ("an apple in general"), and it produced the desire through pleasurable apple-eating, or the pleasurable representations of the existence of an apple that are involved in apple-eating. Hence it was not the simple existence of an apple, but rather an apple's presence to the senses, that is, its existence in relation to my receptivity as a cause affecting the latter – a relation consisting in the pleasurable representations of its existence that figured in the eating of it – that produced the sensible desire, and the effect of this desire is precisely the same thing: not the bare existence of an apple, but the existence of one in that same relation of affection to my sensibility through which the desire was first produced. (The situation is the same in the case of sensible desire that arises from theoretical judgment: the sympathetic philanthropist's immediate inclination to help others arose through the pleasure he found in his recognition of success in his efforts to help others; this pleasing recognition was not, however, merely a recognition of the particular instances of past success in helping this or that particular individual in need, but rather his cognizance of his helping itself, encountered in those instances, and this helping action is precisely the object of his inclination.)

Moreover, the assignment of an essential reproductive character to sensible desire has a notable implication that confirms the correctness of this account. In producing its effect sensible desire reproduces its cause. But in reproducing its cause it reinforces its own production and thereby sustains and reinforces itself. It follows from this that, by its nature, sensible desire is from the beginning at least incipiently habitual. Now habitual sensible desire is just what Kant means when he uses the term 'inclination' (*Neigung*) (MS 212, Anth 251). It turns out, then, that, rather than being paradoxical, the identification of sensible desire's cause with its effect enables us to account for its habitual character and hence to understand how inclination is possible.

But on the other hand, this implication does leave us with another question. For if rational desire is productive rather than reproductive, then our account of it affords no parallel implication that would explain the possibility of habitual desire in the case of rational desire, the possibility of what Kant calls "sense-free inclination (*propensio intellectualis*)" (MS 213). Yet it would seem that there must be such a possibility if such a thing as virtue is to be intelligible. We shall return to this question below.

5. Another concern that may be occasioned by the account of the agreeable sketched above relates to a charge often leveled against Kant's view of the way sensible desire figures as a motive in choice. According to this criticism, Kant's account is objectionably hedonistic in that the only determinant of choice it recognizes, aside from the moral law, is the quantity of pleasure expected to result from the action. Much of the dissatisfaction prompting such criticism stems from Kant's apparent suggestion that pleasure itself is the object to which our thought and attention are directed in sensible desire and in the choices based on it, that the sensible desire for an apple is at bottom a desire for the pleasure we expect the eating of it will bring – a view that even in Kant's day had already been effectively criticized by Butler, Hume, and others. The above characterization of the object of sensible desire as "the cause of gratification" might appear to suggest that the critics are correct at least in thinking that Kant supposes that whenever we sensibly desire something, it is the gratification that we have in view in desiring it. Though it lies beyond the scope of this essay to explore the charge of hedonism,[21] the account of sensible desire and its object developed so far already provides everything needed to remove any appearance there may be that the characterization of the agreeable as necessarily, in virtue of its form, the cause of gratification implies that pleasure necessarily belongs to the content of (or to what is represented in) sensible desire. The following points are particularly pertinent.

Most importantly, we should bear in mind that the form of the agreeable and the form of the good stand in quite different relations to thought and consciousness. Because rational desire, even in respect of its efficacy, is spontaneous and hence self-conscious, the relation of causal dependence in which the good stands to rational desire necessarily belongs to what is represented in the efficacious conception that constitutes rational desire. But since sensible desire is receptive, neither its own efficacy in respect of the agreeable nor the agreeable's essential relation to sensible desire as the latter's cause are necessarily part of what is represented in the efficacious representation in which sensible desire itself consists.

Sensible desire does reflect the gratification through which its object, the agreeable, produced it. For the distinctive feelings that, as we noted earlier, constitute the subject's awareness of sensible desire (feelings of anticipation, fear, frustration, and the like) are simply modes of the same

[21] Recent responses can be found in Reath (1989), and Herman (2001). For statements of the criticism, see Foot (1978, p. 165), and Irwin (1984, pp. 39*f*), which revives criticisms raised by T.H. Green in the nineteenth century.

receptive awareness of the faculty of desire's determination whose first occurrence was the initial feeling of gratification that inaugurated the sensible desire. But while these pleasing and painful feelings to which the subject of sensible desire is liable do reflect the fact that the object of sensible desire produces a feeling of gratification, and while it is also true that, to the extent that the subject reflects upon these feelings, or simply remembers the past gratification, the expectation of gratification may subsequently come to be included in the subject's conception of the object of the sensible desire, it remains true nevertheless that the representation of that gratification does not belong to sensible desire's *original* representation of its object. Just as the original encounter with the agreeable – the first bite of the apple – consisted in a representation of the existence of an object accompanied by a feeling of pleasure, so the sensible desire for that object consists in a representation of it productive of its existence accompanied by modes of feeling that vary according as the production is furthered or hindered.

3. Subsuming the Agreeable Under the Concept of the Good

1. We have seen that rational and sensible desire have different sources, that they have essentially different forms of efficacy, and that they differ essentially with regard to the form of the relation in which their objects stand to them. And we have noted that, in consequence of these facts, the objects of these two sorts of desire differ in form. Goodness, then, does not boil down to agreeableness, nor agreeableness to goodness. But it does not follow from any of these facts, nor from all of them taken together, that the good and the agreeable cannot in any way coincide. Indeed, as we shall see shortly, Kant himself claims that the agreeable can, at least in some cases, be subsumed under the concept of the good. If such subsumption is possible, however, there must be something the agreeable and the good have in common, notwithstanding the significant difference between them to which Kant explicitly draws our attention. The one point of commonality we have noted so far is that they are both objects of the faculty of desire and thus objects of representations that possess an efficacy in respect of them. But appreciating this generic commonality by itself does not seem to take us very far toward understanding how, or in what way, the agreeable can also be good. To consider how, and under what conditions, the agreeable can be subsumed under the concept of the good will be our principal concern in what follows. Exploring this question will reveal that the agreeable and the

good, according to Kant's own understanding of them, are more intimately related than his rhetoric commonly inclines us to suppose.

This much, however, is already clear: Whatever affinity there may prove to be between the good and the agreeable, the mere consideration of the forms of these two objects shows that any judgment in which the agreeable is deemed good will have to be one in and through which the cause of gratification comes to be represented also as the effect of practical cognition and action. It therefore follows directly that this or that object of sensible desire can indeed be good only to the extent that it is subsumed under the concept of the good in a practical judgment representing that object as its effect, and only to the extent that this judgment has the basis in principles of reason and the consequent universal communicability that are characteristic of genuine practical cognition.

2. Kant offers the following description of how the subsumption of the agreeable under the concept of the good is to take place: "The agreeable, which as such represents the object solely in relation to the senses, must first be brought under principles of reason through the concept of an end in order to be, as object of the will, called good" (KU 208). In saying that the agreeable must "be brought under principles of reason" Kant would appear to be expressing the point just noted, namely, that the judgment in which the agreeable is deemed good must have a basis in such principles. For it is through practical cognition's own relation to these principles that its effects stand under them. But he goes on to specify that the agreeable is to be brought under these principles "through the concept of an end." Part of what Kant may intend to convey by this specification is the point that in the judgment through which the agreeable is brought under the concept of the good the agreeable is regarded as the *effect* of that act of practical cognition (a point that, as we have seen, is entailed by the concept of the good as the object of practical cognition). For Kant standardly characterizes ends as the effects of their own representations: "an end is the object of a concept, so far as the latter is regarded as the cause of the former (the real ground of its possibility)" (KU 220; cf. MS 384). But the same point would have been entailed had Kant said instead that the agreeable must be brought under principles of reason "through the concept of the *means* to an end," for to practically cognize something as the means to an end is equally to regard it as the effect of that act of practical cognition. It would appear, then, that in specifying that the agreeable is to be brought under principles of reason through the concept of an end Kant means to convey more than the general point that in being brought under the concept of the good the agreeable comes to be

regarded as the effect of practical cognition. Further consideration of the concept of an end can therefore be expected to help clarify the relation between the agreeable and the good. We should first consider how this concept is related to that of the good, and then ask how and under what conditions the agreeable can be subsumed under it.

One of the observations Kant makes in contrasting the good and the agreeable is that while it makes no sense to speak of something as *mediately* agreeable,[22] we recognize a difference between what is mediately and what is immediately good, as when we distinguish between what is useful and what is good in itself, or between what is good as a means and what is good as an end (KU 208; cf. G 414). Informing this observation is the idea that the concepts of end and means are complementary concepts that divide the concept of the good, yielding two forms of goodness, or two ways in which something can be good. The former signifies what is represented in practical cognition as being for its own sake and the latter what is represented in such cognition as being for the sake of something other than itself. But the relation of being-for-the-sake-of is just the relation of efficacy practically represented. Thus, for something to be deemed good as a means, or for the sake of something else, is for it to be represented in a practical judgment as something that furthers something else or somehow contributes to its furtherance.[23] And for something to be deemed good as

[22] Tools, equipment, and the like do not count as agreeable just because their use makes possible some enjoyable activity, though the awareness that they are available for such use may be very pleasing.

[23] The concept of means may be applied either to action itself (G 414) or to what is used in action (what is practically represented as a material condition of the possibility of the action) (G 427). If riding a bicycle is a means of traveling between home and work, then in another sense the same is true of the bicycle itself. In the former case we might speak of the means as formal, in the latter as material. Where the means are material, the efficacy too can be described as material rather than formal in character, or in other words as the efficacy to be found among the objects of practical knowledge rather than the efficacy proper to practical knowledge itself and by which the latter makes its object actual. (Practical knowledge is not itself *for the sake of* its object, though the action whereby it produces the object may be said to be.) Though these types of efficacy are distinct, they are closely related. For, on the one hand, material efficacy depends on and is indeed constituted by formal efficacy: it is only because of the formal efficacy of practical knowledge – i.e., only because of the *action* issuing from such knowledge – that the material relations of efficacy represented in that cognition have any reality. We say our coats keep us warm, for example, but they can do this only because we can maintain and wear them – only, that is, because we can use them to keep ourselves warm. And on the other hand, such action would obviously not be possible were there not antecedently given (available for use) an object suitably constituted to receive the material efficacy with

an end, or for its own sake, is for it to be represented in a practical judgment as something that furthers itself. It therefore belongs to the concept of an end that an end (whether of a person or of any other living thing) is in necessary (nonaccidental) agreement with itself, that all the elements that may belong to it are in systematic harmony with one another, so that, each of these elements thus furthering and so being furthered by all (and so also being related as a means to the very end of which it is a part), there is nothing internal to the end that could bring it to an end or even impede it in any way at all.[24] So insofar as it exists, and to the extent that external conditions allow, an end always *sustains itself*.

3. Let us turn now to the concept of the agreeable. Does the account developed so far enable us to determine whether the agreeable shares the self-sustaining character implied by the idea of an end?

As we have seen, when Kant speaks of the agreeable he standardly has in view the *object* of sensible desire. Sensible qualities (such as colors and sounds), a spicy dish, and wine from the Canaries are among the examples he offers of things this or that person might find agreeable. Clearly these objects on their own do not necessarily have the form characteristic of an end; no spicy dish sustains itself, nor does a sweet canary, though it may improve with age. But if they are considered in respect of the causal power in virtue of possessing which they count as agreeable objects in the first place, if they are considered, that is, in relation to the pleasurable state of mind in which their existence is represented (the enjoyment of them) and through which they are causes of sensible desire, then on account of the reproductive character of

---

which the action endows it. As we shall see below (§5), there is a parallel interdependency in the case of ends.

[24] The elements in question here are to be understood as related to one another as parts, or members, not as form and matter. Thus in the case of the highest good (the ultimate end for persons), which according to Kant consists in universal happiness collectively consequent upon universal virtue, the elements that in this end stand to one another in the relation of mutual furtherance just described are its members, namely, persons. But insofar as the highest good is viewed as composed of virtue and the resulting achievement of happiness, related as form and matter, the furtherance among the components is not reciprocal: virtue produces happiness, not happiness virtue. This asymmetry is obviously due to the fact that virtue is the *form* of the highest good, the form that constitutes it as an end: it is because virtue is shared by all the members that the latter, through their actions, mutually further one another's existence and happiness. (Cf. KrV A812/B839: "Morality in itself constitutes a system, but happiness does not, except insofar as it is distributed in exact proportion to morality.") And because virtue is the form of this end, it is not a *means* of achieving happiness, despite the fact that it produces it. Being itself an end, virtue furthers itself; and lying in practical (efficacious) knowledge, it also produces its object, the highest good, of which it is itself the form.

sensible desire each of them can be recognized as an element of a self-furthering system: the gratifying representation of existence depends on the agreeable object, and the agreeable object depends on the sensible desire that arises through that gratifying representation.

Moreover, even though a spicy dish is not self-sustaining on its own (any more than any other artifact or product of human contrivance is self-sustaining), the pleasing representation of its existence does contain wholly within itself a self-sustaining tendency. This tendency is explicitly pointed out in Kant's general account of pleasure, according to which pleasure consists precisely in the feeling of a state of mind's sustaining itself. In the *Critique of Judgment*, he says, "The consciousness of the causality of a representation in respect of the state of the subject, *to maintain* it in the same, can here characterize in general what one calls pleasure" (KU 220).[25] If such consciousness is involved in a representation the subject is capable of producing spontaneously, the representation's maintenance of its state of mind may lie simply in its holding the subject's attention. Such self-maintenance, Kant claims, is characteristic of the delight we take in the beautiful (KU 222). In the case of pleasure in the representation of the agreeable, however, the representation is of the *existence* of the agreeable and therefore depends essentially on external conditions, namely, the existence of its object in relation to the capacity to represent the existence of things (i.e., its presence to the outer senses or to the subject's theoretical cognitive capacity). Thus, such a representation's maintenance of its state of mind happens not only directly and inwardly, through its holding the subject's

[25] The capacity to feel pleasure and displeasure in general is the capacity to be conscious of the causality of representations with respect to the state of the subject. Where the consciousness is of the causality of a representation to maintain the state of the subject, it is the feeling of pleasure; where the consciousness is of the state of mind's causality to determine itself to the opposite state, it is the feeling of displeasure. This characterization of pleasure as the *consciousness* of a representation's causality should not be understood as implying that pleasure itself is a type of cognition or involves the application of concepts (e.g., of cause, or representation), for as a type of feeling, pleasure belongs to receptivity, not spontaneity; the consciousness in which pleasure is said to consist is better regarded as the sensible indication, or inward manifestation, of the self-sustaining character of a state of mind. In the *Critique of Practical Reason* Kant offers a somewhat different definition of pleasure: "the *representation of the agreement of the object or the action with the* **subjective** *conditions of life*, that is, with the capacity of the *causality of a representation in respect of the actuality of its object* (or the determination of the powers of the subject to the action of producing it)" (KpV 9n). This definition differs from the one given in the *Critique of Judgment* mainly in that it concerns a specific *type* of pleasure, what Kant calls *practical* pleasure, or pleasure necessarily connected with desire (MS 212). This type of pleasure will be considered below (§4).

attention, but also through its determining the faculty of desire to produce action that sustains the external conditions (the existence of the object in relation to the capacity to represent the existence of things) on which the self-sustaining gratifying state of mind depends. The enjoyment of the spicy dish thus leads to another spoonful, say, or another serving, or to the placement of more spices in the cupboard to be used in preparing it again. In the case of pleasure in the agreeable, then, the directly self-sustaining causality of a pleasing representation of existence is also the basis of a self-furthering system involving not only the subject's capacity to *represent* the existence of things, but also its faculty of desire, or its capacity to *produce*, or to bring into existence, things it represents.

In spelling out his account of pleasure as it figures in the specific case of delight in the beautiful, Kant explains the self-sustaining character of the pleasing state of mind by appealing to a harmony between the faculties that are called into play in the representation of the beautiful. As we shall see below, it is possible to extend this explanation to the other types of pleasure as well. Doing so will enable us to trace the source of the self-sustaining character of representations of the existence of the agreeable back to a harmony between the faculty of representing the existence of things and the faculty of desire. It will also enable us to understand how rational desire can be habitual and thus how virtue is possible, and finally it will put us in a better position to identify the relation between the agreeable and the good.

4. Pleasure

1. We have noted that Kant defines pleasure in general as the consciousness of the causality whereby a representation tends to sustain its own state of mind. And we have integrated this view of pleasure into our elaboration of his account of the agreeable. We have observed that the agreeable is an object the representation of whose existence is pleasing (where this representation may be a sensation or a theoretical judgment), and that this pleasure is just the awareness of the representation's determination of the faculty of desire, through which there arises a (sensible) desire for that object, or a representation having an efficacy that constitutes a dependence of the object's actuality on the actuality of the representation (in that, if the conditions are right, the representation leads to action that produces its object). But since, as we have also observed, the actuality of the pleasing representation of the

existence of the agreeable depends on the actuality of its object, this representation's determination of the faculty of desire is at the same time a tendency, or "causality," to make actual the very object on which its own actuality depends and hence is a tendency to sustain itself and the state of mind to which it belongs. Thus the awareness of the representation's determination of the faculty of desire is an awareness of a self-sustaining tendency (in its outward aspect), which is just what pleasure is, according to Kant's definition.

But this characterization of pleasure as the awareness of a representation's self-sustaining tendency is not the end of the story. In the *Critique of Judgment*, Kant's account of the pleasure taken in the beautiful extends to a deeper level. This account relies on the idea of a harmony among certain faculties of the mind, a harmony that constitutes the self-sustaining causality of which we are conscious in the feeling of such pleasure. More specifically, the pleasure in which a judgment of the beautiful consists is said to lie in an awareness of a harmony of "free play" in the exercise of certain faculties, a harmony that consists in the mutual furtherance of their activities.[26]

What is of particular interest for our purposes here is that Kant's idea that pleasure in the beautiful is based in a harmony among certain faculties can be seen to be a special case of a general understanding of pleasure, according to which such a harmony is basic to pleasure across the board, not only to the merely "contemplative pleasure" that constitutes the disinterested delight in beauty, but also to "practical pleasure," or pleasure "necessarily connected with desire" (MS 212), the type of pleasure relevant to our present discussion. Appreciating this account in its full generality, and elaborating it to accommodate both the difference between contemplative pleasure and practical pleasure and also a further difference, to be pointed out below, between two types of practical pleasure, will contribute to our understanding of rational and sensible desire and thereby throw additional light on how their objects are related. It will suffice for the task at hand if we can trace the account in its main outlines.

2. Contemplative and practical pleasure differ, not in respect of the quality of the feeling in which they consist (for in this regard all

[26] This harmony of free play depends in turn on a deeper harmony between the faculties themselves, a harmony that constitutes taste, or the capacity to judge the beautiful, which Kant also identifies as a type of "common sense." Kant's idea of a harmony of the faculties has often been criticized, especially by readers averse to talk of mental faculties. While it does not belong to our aim here to review the criticisms, the considerations to follow may, if correct, provide some support for this idea.

pleasures are the same), but in virtue of a difference among the representations (or aspects of representation) in which the felt efficacy is located. Contemplative pleasure lies in the awareness of the self-sustaining tendency of a representation of sensible (spatiotemporal) *form*, which is always spontaneous. Practical pleasure, on the other hand, lies in the awareness of the self-sustaining tendency of a representation of the *actuality* (i.e., existence) of an object, a representation that always depends on receptivity. But in both cases the pleasure lies in an awareness of a representation's tendency to sustain its own state of mind, and in both cases this tendency is nothing other than the harmony, or mutual furtherance, of the exercise of distinct mental powers engaged in that representation.

In the case of contemplative pleasure, Kant claims, the harmonizing faculties are the understanding and imagination in their theoretical, or rather incipiently theoretical, employment. According to Kant's analysis of the theoretical cognition of an object of experience, the understanding determines the imagination, in accordance with its concept of the object, to represent in intuition the (spatiotemporal) form of the object cognized. Where such an object is appreciated as beautiful, however, the representation of its form through the exercise of the imagination is not determined by any concept of the understanding, yet nevertheless is so related to the activity of reflection whereby the understanding arrives at a concept of the object through which the object can be known that each of these two activities – the representation of form on the part of imagination and the reflection on that form by the understanding – stimulates and reinforces the other rather than disturbing or interfering with it. The receptive awareness of this self-maintenance of a state of mind through the reciprocal furtherance of its component activities is what contemplative pleasure consists in. (And because the state of mind is incipiently theoretically cognitive in character, Kant holds, the pleasure is valid universally, for all human cognizing subjects.)

In the case of practical pleasure, on the other hand, the relation of mutual furtherance cannot lie between two faculties that belong to the power of theoretical cognition and cooperate in such cognition, for practical pleasure is "necessarily connected with desire" and so must involve the faculty of desire. The mutual furtherance is rather between the two broad powers of the mind marked out by the different directions of existential dependence in which representation and its object can stand to one another (§1.1): on the one side is the subject's capacity to represent the existence of things ("what is"), and on the other is its faculty of desire, or its capacity to produce, or bring into existence,

objects through its representations of them. For ease of reference, let us call these faculties and the representations belonging to them "theoretical" and "desiderative" respectively. In practical pleasure, then, a theoretical representation and a desiderative representation (or desire) share the same object, and each representation furthers the other. The desiderative representation of the object, through its efficacy, tends to sustain the actuality of the object in relation to the theoretical faculty, and thereby furthers the theoretical representation of the object, since such representation depends for its actuality on the actuality of that object. And conversely, in practical pleasure the theoretical representation of the object tends to sustain the desiderative representation of it. We have already seen how this happens in the case of gratification, or the pleasure tied to sensible desire, where it is the theoretical representation that first gives rise to the desiderative representation: it was from the taste of the apple that there arose the pleasure in it and the sensible desire that has that agreeable taste as its object, and it is through the repeated experience of this fruit that the inclination and enjoyment are sustained. But as we shall presently see, this is not the only way in which theoretical representation can sustain desiderative.

3. There are two types of practical pleasure, corresponding to the two kinds of desiderative representation we took note of earlier when we distinguished between rational and sensible desire (§1.2-3). Just as the basis of the difference between these two forms of desire lies, as we saw, in a difference in their origin, so the difference between the two types of practical pleasure is grounded in the different ways they arise. Practical pleasure always springs immediately from a theoretical representation of existence, of course, for such pleasure is just the receptive awareness of the self-sustaining efficacy of such a representation. But there is this difference in the way the two sorts of practical pleasure are brought about: the type connected with sensible desire arises from such theoretical representations *alone*, whereas the type connected with rational desire arises through theoretical cognitions that are made possible by the efficacy of such desire. The two types can therefore be distinguished by saying that the one corresponding to rational desire is engendered practically and *a priori* (i.e., through action from desire based in principles of practical cognition), and the one corresponding to sensible desire is initiated theoretically and *a posteriori*. The first of these Kant characterizes variously by speaking of feelings of self-contentment, self-respect, approval, and esteem; the second, as we have noted, he calls gratification (KpV 117-118, 161, KU 210). He also sug-

gests that they might be called, respectively, "intellectual" and "sensible" practical pleasure (MS 212-213).

The case of *theoretically* initiated practical pleasure has already come under consideration and was the focus in our earlier discussion of the self-sustaining character of a gratifying representation of the existence of the agreeable (§3.3). Here the mutual furtherance is occasioned when a theoretical representation of the existence of an object first brings about the desiderative representation of that object. To say that the latter representation arises in this way is just to say that the subject's awareness of the efficacy constituting it as a desire is consequent upon, rather than constitutive of, this efficacy. And as we have seen, a desiderative representation that comes to be in this way and whose efficacy bears this relation to the awareness of it is a sensible desire (§1.2-3). So in this case a theoretical representation gives rise to a sensible desire, which in turn tends, through its efficacy, to bring into existence the object on which the theoretical representation depends.

We noted earlier (§1.2) that Kant identifies the feeling of gratification with the awareness of a theoretical representation's determination of the faculty of desire. This identification involved no explicit suggestion that what this feeling manifests is in fact a reciprocal rather than a unilateral relation between the theoretical representation and the resulting sensible desire. We now are in a position from which, by drawing on the general account of pleasure presently under consideration, we can confirm this identification and indeed provide a basis for it by saying that since the sensible desire that arises through the determination of the faculty of desire is reproductive in character, or efficaciously directed to produce the object of the very theoretical representation that produced it, that same determination is at the same time a *reciprocal* furtherance between the exercise of the theoretical faculty and that of the desiderative faculty and *for this reason* something of which the subject is receptively aware in a feeling of *pleasure*.

A further distinction can be made between different classes of sensible practical pleasure, according as the theoretical representation of existence involved is a sensation or a theoretical judgment (cf. §1.2). Feelings of pleasure connected with sensation might be placed under the heading "pleasures of the senses," and those tied to theoretical cognition might be arranged under the headings "pleasures of the mind" (invention, discovery, learning, and so forth) and "pleasures of society" (friendship, cooperation, competition, and the like). But since the differences between these two classes of sensible practical pleasure are not essential to our present purpose, we need not consider them further here.

In the case of *practically* engendered practical pleasure, on the other hand, the mutual furtherance is initiated by the desiderative representation of the object, which in this case lies in the practical knowledge of it, and which through its efficacy makes the object actual and thereby makes possible the theoretical representation (theoretical cognition) of the object's existence. But how does the theoretical representation in turn further the practical? Since the practical representation is *a priori* and hence spontaneous, it cannot itself arise in the way sensible desire does, as a result of the faculty of desire's being affected by a theoretical representation. By representing the *actuality* of the object of practical representation, however, the theoretical cognition does confirm the presupposed cognition of the *possibility* of that object, on which practical cognition depends for its possibility. As was noted earlier, it is characteristic of rational desire that its efficacy is constituted by its own awareness of that efficacy. But this awareness lies in rational desire's understanding of itself as the cause of the object it represents, which is just its understanding of the object as its possible effect. This practical understanding that rational desire has of its relation to its own object can be furthered or undermined by theoretical judgments bearing on the possibility of the production of that object. We have already observed that cognizance of actual failure in one's past attempts tends to weaken confidence in one's capacity to produce the object and thereby tends to make it more difficult to choose to try again. But just as a subject's theoretical judgment that a certain object lies beyond its power to produce effects will undermine any rational desire the subject may previously have had for that object, so a theoretical judgment that such an object has been produced through the efficacy of rational desire will reinforce rational desire's essential practical understanding of itself as the cause of its object and thereby reinforce rational desire itself. Thus, each representation, practical and theoretical, furthers the other, and this positive interaction constitutes the harmony.

4. In the preceding comparison, the two types of practical pleasure have been contrasted in respect of the different ways in which they arise. But because practical pleasure is "necessarily connected with desire," it is also possible, and useful, to contrast its two types by reference to the different ways in which they are connected with desire. Thus, another way of expressing the difference between them – one that Kant explicitly points out (MS 212), and one that is clearly implicit in what has already been said – is by saying that in the case of theoretically initiated pleasure, the pleasure necessarily precedes and brings about the exercise of the faculty of desire, whereas in the case of practically engendered

pleasure, the pleasure follows upon that faculty's exercise. For while there is a sense in which both sorts of pleasure can be said to "follow upon" the exercise of the faculty of desire in that all practical pleasure is receptive awareness of a state of mind involving a desire, the two sorts of pleasure nevertheless differ in that in the one case the pleasure necessarily precedes, and in the other it necessarily follows, the desire's *production* of the object. And since whether the pleasure precedes or follows is just a matter of whether it is the lower or the higher faculty of desire whose exercise is in question, the difference can also be expressed by saying that in the one case it is sensible desire, in the other rational desire, that stands in a relation of mutual furtherance with a theoretical representation.

The "necessary connection" Kant says practical pleasure has with desire can also be described as the relation of actuality to potentiality. For the pleasure always accompanies the representation of the existence, or *actuality*, of the very object that is represented in the desire with which it is necessarily connected, and since the desire is simply the determination of the subject's faculty of desire, or power through whose determination the subject is a cause of the existence of the object represented in that determination, the subject has, in this desire, the power, or *potentiality*, to produce that object. Therefore the different directions of dependence just noted between practical pleasure and desire that distinguish the two types of practical pleasure also amount to different dependence relations between practical actuality and practical potentiality. In the one case (where the pleasure precedes the desire's production of the object) the potentiality in which the desiderative representation consists comes to be only through the actuality (i.e., only through the pleasing state of mind) and likewise can be known only through the awareness of its actuality (i.e., *a posteriori*, from the feeling of pleasure or displeasure), whereas in the other case the actuality comes to be only through the potentiality and likewise can be known only through knowledge of the potentiality (i.e., *a priori*, from the self-consciousness of practical cognition). Since in both of these cases the potentiality is in Aristotelian terms in fact "first actuality" (not bare potentiality, or the faculty of desire itself) (Aristotle *De Anima*, 412a22-23, 417a21-b2), the contrast can also be expressed by saying that in the one case first actuality arises through second, whereas in the other second arises through first. The distinction between these two types of first actuality finds expression in Kant's terminology principally as the distinction between inclinations (habitual sensible desires) and maxims (practical principles constituting a person's character).

5. We are now in a position briefly to address the question raised earlier (§2.4) relating to the possibility of virtue, namely, whether and if so how it is possible for rational desire to be habitual, given that it does not have the reproductive character that makes intelligible the habitual nature of sensible desire. Now the fact that rational desire is first actuality that precedes second already affords a sense in which such desire is habitual. Maxims, after all, are not momentarily occurring states of mind; in virtue of being, at least in purport, principles of practical cognition, they are rather abiding determinations of a person's will and as such constituents of character. But the sense of 'habit' that figures both in traditional definitions of virtue in terms of habit and in the qualified version of this type of definition endorsed by Kant (MS 383-384, 407) implies a certain strength and readiness to act in the virtuous person's maxim and character, a strength and readiness acquired in part through practice and hence through action directed toward the actualization of the object represented in the maxim. Thus, it is the account of practically engendered practical pleasure just outlined (§4.3) that reveals the basis of the possibility of the distinctive form of habitual desire in which virtue consists, just as the mutual furtherance involved in theoretically initiated practical pleasure lies at the basis of the habitual nature of sensible desire. Even though rational desire is essentially productive rather than reproductive, its own efficacy leads to a reciprocal interaction between the practical and theoretical capacities through which rational desire is itself furthered. Since its effect, while not indeed what first brings it into existence, is nevertheless the cause of its own strengthening, rational desire reinforces itself. To the extent that rational desire comes to be thus spontaneously strengthened, it qualifies as "sense-free inclination (*propensio intellectualis*)."

A full examination of Kant's account of how virtue is possible would of course require substantial further discussion. In particular, it would be necessary to consider his explanation of how the strength and readiness of virtuous maxims can be acquired despite the fact that the lower faculty of desire generates not only natural sensible desires that may occasionally conflict with practical reason but also propensities of self-love that stand in essential opposition to it. Yet while there is much in this explanation that could be usefully related to our foregoing discussion of practical pleasure (most notably its account of respect for the moral law, a complex mode of feeling to which both pleasure and displeasure are essential: cf. KpV 71*ff*), our limited purposes do not require us to explore this topic here.

5. The Agreeable as the External Condition of the Sensible Material of the Good

1. Let us return now to our main question, concerning how and under what condition the agreeable can be subsumed under the concept of the good. We have already observed (§3.1) that the difference in form between the objects of rational and sensible desire entails that the agreeable cannot be subsumed under this concept except through receiving a new form. The good, as the object of practical cognition, is essentially conceived as such cognition's own product, whereas no such conception is included in the bare representation of the agreeable, in which the object is viewed merely as the cause or occasioning condition of gratification and thereby of sensible desire. The subsumption of the agreeable under the concept of the good will therefore have to be one in and through which the cause of gratification comes to be represented also as the effect of practical cognition and action.

On the other hand, we have also seen that the agreeable is bound up as an element in a self-furthering system the basis of which lies in a relation of mutual furtherance between the exercise of the faculty of representation of things' existence and that of the faculty of desire, and that this mutual furtherance in the exercise of distinct mental powers conforms to what is represented in the idea of an end, even though it is something of which the subject is aware merely through feeling, rather than in a practical cognition of it as something to be produced for its own sake. The agreeable is bound up with the reciprocal exercise of these powers as the object their representations share in common, the one representation looking backward, so to speak, to its cause or occasioning condition, the other forward to its effect.

In short, we have found that while the agreeable differs essentially from the good, it is nevertheless the object of a system of representation that, being self-sustaining, shares with one type of goodness (that of an end) the quality that differentiates the latter from the other type (that of the means). But since the representations figuring in this system are receptive, their mutually furthering actuality depends on the existence of their object. So to the extent that this self-sustaining system can itself be brought under the concept of an end and thereby become an object of practical knowledge, the agreeable will also be included in what is represented in that cognition; for the cognition represents its own production, or actualization, of that system, and this production must be through the production of its object. It remains to spell out this role the agreeable plays in this end.

2. As the object of a self-sustaining system of receptive, or empirical, representation, the agreeable can be characterized as the system's external condition, on which its mutually furthering constituent representations depend. Though the agreeable is something the enjoyment of which will be brought about, if conditions allow, through the efficacy of sensible desire, it remains nevertheless true that did this object not antecedently exist, as an object of the senses or of theoretical cognition, there would be no representation of its existence, no gratification, and no sensible desire for it.

The subject's awareness, in the feeling of gratification, of the mutual furtherance between the representation of the object's existence and the associated sensible desire is a sensible mark of a state of mind that, as self-sustaining, has the character distinctive of an end (even though as a mere feeling of gratification it is not itself practical cognition and hence not a representation of anything good). Indeed, as practical pleasure, gratification is identical in form with the sensible awareness of the mutual furtherance between rational desire and theoretical cognition that constitutes the practically engendered practical pleasure in the good. The self-furthering state of mind inwardly manifest in the feeling of gratification thus provides the sensible material suitable for application of the concept of an end.[27] And the agreeable is this material's external condition.

In being brought under the concept of an end in the act of practical judgment, the gratifying state of mind is brought into relation to practical cognition as the latter's effect. This is not to say, of course, that this state of mind is no longer regarded as dependent for its actuality on its agreeable object. The representation of this object's existence that belongs to the gratifying state of mind is essentially dependent for its own existence on the existence of that external object. As we have noted, the original conception of the agreeable is simply the conception of the external condition on which that representation depends. Thus, in the act of subsuming the gratifying state of mind under the concept of an end, that state of mind is represented only as the *mediate* effect of practical cognition: the production of the gratifying state of mind must be represented as production *through* the production of its object, the agreeable.

It might at first glance appear that to say that the gratifying state of mind is produced *through* the production of its object is to say that the

[27] Accordingly, *gratification* provides the sensible material for applying the concept of the *good* in *practical* knowledge, just as *sensation* provides the sensible material for applying the concept of *reality* in *theoretical* knowledge (cf. KrV A143/B182).

production of the latter is a *means* to achieving the former, and this appearance may (assuming the correctness of the interpretation here presented) occasion the impression that, despite Kant's explicit statement that it is through the concept of an *end* that the agreeable is subsumed under the concept of the good, an objectionable variety of hedonism nevertheless emerges in his account of rational desire and the good, if not in his view of sensible desire and the agreeable (cf. §2.5 above). But the designation of some element of the object of practical cognition as a means depends on the possibility of separating it off from the end in practical thought, which in turn depends on at least the bare conceivability of some alternative means that might stand in the same relation to the end. The very idea of a gratifying state of mind, however, already includes the idea of the agreeable, as the object, the external condition, on whose existence its own existence, as that of a doubly receptive state of mind, depends. Judgments as to what, in a given case, that agreeable object might be can of course be in error and are open to revision and extension. One might be ignorant or mistaken, for instance, about which spices are responsible for the delightful flavors one is enjoying, or erroneously believe that the fruit one is eating is a pear, or uncertain on reflection whether it is the recognition that one is being helpful to others that pleases, or rather the awareness that one is being helpful to others who have a certain relation to oneself. Indeed, nothing is more common than for the representation of the object to have a specificity beyond those elements belonging to it that originally (prior to habituation) produce or contribute to its gratifying determination of the faculty of desire, a specificity that makes possible a certain plasticity in sensible desire, its habitual character notwithstanding. As Hume observes, "A man, who has contracted a custom of eating fruit by the use of pears or peaches, will satisfy himself with melons, where he cannot find his favourite fruit; as one, who has become a drunkard by the use of red wines, will be carried almost with the same violence to white, if presented to him" (Hume 1739, Book I, Part 3, Sect. 13).[28] Not to mention that those in search of gratification often turn to wholly new objects, diversions, and adventures, thinking they may find in them new

[28] It is worth noting here that anyone who supposes that an objectionable hedonism lurks in Kant's idea that the representation in practical cognition of the production of a gratifying state of mind is a representation of its production *through* the production of its object should also be inclined to think that an objectionable hedonism is betrayed in the perfectly ordinary way in which Hume speaks of the *use* of pears, peaches, and red wines. It is plain, however, that this ordinary usage is completely innocent; Hume could just as well have used 'enjoyment' (cf. note 18, above).

sources of pleasure, "just as to someone who wants money to spend it is all the same whether the material in it, the gold, was dug from the mountain or washed from the sand" (KpV 23). But it is no more possible to separate the bare notion of the agreeable itself from that of the gratifying state of mind depending on it than it is possible to separate the concept of cause from that of effect.

Since, therefore, in the subsumption of a gratifying state of mind under the concept of an end practical cognition *necessarily* represents its own efficacy in respect of that state of mind as mediated by the latter's agreeable object, it belongs to the *form* of practical cognition's representation of the action through which the end is to be produced that it is a representation of this action as the production of the agreeable. (The content, on the other hand, lies in the specific representation of the action by or through which the specific agreeable object – an apple, a spicy dish, one's generosity, etc. – is to be enjoyed.) Thus, the agreeable can be subsumed under the concept of an end and thereby also under that of the good, and moreover is necessarily brought under these concepts to the extent that the gratifying state of mind that has it as its object is brought under them; but in this subsumption the agreeable is represented as belonging to the end only as the external condition of the sensible material of the good – only, that is, by courtesy of its necessary relation to the gratifying state of mind that has it as its object.

3. This state of mind – the enjoyment of the agreeable object – can, in turn, be good only under the condition that it can be made into an object of practical cognition and thereby brought into a relation of dependence on that cognition. As we have seen (§2.2), it belongs to the form of the object of practical knowledge that this object is the effect of that knowledge, since this is how it is necessarily represented in practical cognition on account of the self-conscious character of the latter's efficacy. As practical, practical cognition is essentially a representation of its own production of the object; as rational, it is essentially a representation of such production in accordance with a principle of practical cognition, or the idea of a practical law. Thus, the enjoyment of an agreeable object can be made into an object of practical cognition to the extent that it can receive a new form through the subject's representing itself as acting to produce this agreeable object in accordance with the idea of a practical law. This object of practical cognition is accordingly a composite whose matter lies in the gratifying state of mind (the sensible material of the good) and whose form lies in the action, in accordance with the idea of a practical law, by which that state of mind is to be produced, which is just the action of producing its

external condition, the agreeable. Whether a form/matter composite represented in a practical judgment is good depends on whether the representation of the essential formal element in it (i.e., the representation of the action, the production), conforms to the idea of a practical law, expressed in the moral law. If the representation of the action does in fact have this form, then the form/matter composite is good, and this amounts to saying that the *informed* matter – i.e., the gratifying state of mind arising from the agreeable *so far as* this is to be attained through the represented action – is good. In other words, the agreeable and the gratifying state of mind depending on it are good only insofar as they are the *product* of practical knowledge. We cannot determine whether the agreeable and its enjoyment are good in detachment from this relation to their cause. Just as for an object to be agreeable is for it to be related to a gratifying state of mind as cause (or occasioning condition) to its possible effect, so for an agreeable object to be good is for it to be related to practical cognition as possible effect to its cause.

To express the idea that the subsumption of the agreeable under the concept of the good involves a formal determination of the sensible material, we can say that such subsumption is a *practical* subsumption, or one in which the object of sensible desire is given a certain form through being represented as the product of practical cognition. This idea is clearly expressed in Kant's account of the happiness he says belongs to the highest good as one of its components. The happiness included in the highest good is not bare happiness – happiness as the mere totality of the object of sensible desire (an "ideal of the imagination," as Kant calls it in the *Groundwork*) – but rather happiness so far as it belongs to the totality of the object of rational desire, or practical reason. Happiness is not by itself good; it is good just insofar as it is brought about through virtuous action. The enjoyable activities in which happiness consists are, in the practical cognition of them as good, limited through the limitation the principles of such cognition place upon the pursuit of them.

For our purposes here it will not be necessary to consider how, according to Kant, these principles effect this limitation. It will suffice to note that he sees them as determining the exercise of the power of choice in a way that both restricts and extends it. On the one hand, they restrict both the act of wish, whereby an agreeable object is first made into an end, and also the act of choice, whereby the action of producing the end is specified in the light of the subject's cognizance of its productive capacities; among the effects of this double limitation are the exclusion from practical reason's ends of the objects of the passions (e.g., the

gratifying maintenance of the upper hand over others) and the exclusion from its actions of all pursuit of agreeable ends that would not be compatible with a like pursuit of such ends by others (e.g., assaults on the freedom and property of others). On the other hand, the limitation also has the effect of extending both the exercise of wish (e.g., to embrace others' ends as well as one's own) and that of choice (e.g., to include maxims of assistance to others in need).

Thus, practical subsumption limits and shapes the sensible material it brings under its concept, so that the object of sensible desire is no longer viewed simply as the cause of sensible desire, but also as something to be brought about (i.e. enjoyed) in accordance with the rational principles of practical cognition. In this way, the sensible object receives a rational form, the intelligible form of a practically cognizable object, through which it first counts as something good.

4. Yet if sensible material is rationally determined in practical subsumption, it is equally true that the pleasing mental activities connected with sensible desire and the agreeable things that are their external conditions are together the necessary material conditions of the good for finite practical subjects; without these, the rational ideas of an end and of the good would be empty forms, practical thoughts without content, even though of themselves practical. Kant's recognition of this dependence of the object of rational desire on the object of sensible desire receives expression in his statement that happiness, the complete object of sensible desire, must have a place in the highest good, the complete object of rational desire. Virtue alone, the unconditioned good, does not yield the complete good; happiness, the conditioned good, must also be included; indeed, even the unconditioned good is described in relation to the object of sensible desire when Kant characterizes virtue as "the worthiness to be happy" (KpV 110). This conception of the highest good as comprising the conditioned as well as the unconditioned good depends on the idea that the conditioned good is limited or determined by the unconditioned good (just as informed material is conditioned or limited by its form), and the intelligibility of this relation of determination depends, in turn, on a conception of a material basis of the determined or conditioned good – a conception of what stands to this good as the determinable stands to the determined. This presupposed conception of material is what sensible desire and its representation of the agreeable provide.

If sensible desire and the agreeable stand in this relation to rational desire and the good, then the relation between the imperatives of practical reason and the inclinations, as Kant conceives of it, is much

more intimate than some of his language may seem to suggest. At the same time, this relation is not of a sort that renders inappropriate the emphasis he places on their distinctness. What can be more distinct, after all, than form and matter? And on further reflection it is not surprising that Kant should conceive of rational and sensible desire as standing in such a relation, given his general view of the human intellect as discursive and hence in need of sensible representation to achieve determinate cognition. Indeed, the parallel in this regard between his accounts of the practical and the theoretical employments of reason is not difficult to see. Just as the concepts of the theoretical intellect can figure in theoretical cognition only insofar as they can be applied to sensible representations in empirical intuition, so the ideas of practical reason can figure in practical cognition only insofar as they can be applied to sensible representations in the enjoyment of the agreeable.[29] And just as theoretical cognition, in representing "what is," represents a world of sensible objects interacting according to reason's theoretical laws of nature, so practical cognition, in representing "what ought to be," or the highest good, represents a world in which rational yet sensible subjects secure their own happiness by acting together according to reason's practical law of freedom.

6. Appendix: "The Universal Wish of Every Rational Being"

1. If the pleasing mental activities connected with sensible desire constitute the necessary material conditions of the good, what then are we to make of the seemingly disparaging remark about the inclinations that we took note of at the outset?

> All objects of the inclinations have only a conditioned worth; for if the inclinations and the needs grounded on them did not exist, their object would be without worth. But the inclinations themselves, as sources of needs, are so far from having an absolute worth making them worthy of

[29] This parallel is subject to one important qualification. Kant holds that the fundamental law of pure practical reason is a *formal* yet synthetic *a priori* practical proposition, in which pure reason constitutes itself as practical (KpV 31-32, 42). This first act of practical cognition – the "fact of reason" – is the determination of the "purely formal" law of causality that constitutes the higher faculty of desire (cf. KpV 22-25) and thereby also determines the will as free, constituting it as a self-determining power. Since this act first constitutes the will as a *faculty* of practical cognition, we might call it a *formal* practical cognition. Strictly speaking, then, it is *material* practical cognition – that is, the *exercise* of this faculty in the cognition of some *object* to be realized through that very cognition – that depends on sensible representations for its matter.

> being wished for that to be completely free [*gänzlich . . . frei*] from them must be the universal wish [*Wunsch*] of every rational being. (G 428)

In view of what Kant says elsewhere about complete freedom (or independence) from the inclinations (e.g., at KpV 84, 118), it seems reasonable to suppose that in speaking here of a wish to be *completely* free from the inclinations he has in mind a type of independence whose possibility would depend on our having a different nature altogether, that of a perfectly rational being, whose will and power of choice would be naturally so constituted as to preclude the possibility that sensible desire might in any way interfere with their determination by principles of practical cognition, and for whom the moral law would therefore not be encountered as an imperative. If we interpret the passage in this way, we will in effect be reading Kant as endorsing, at least for the case of rational beings, the view held by Leibniz and the rational moral perfectionists of his day, and indeed by more or less the whole Platonic tradition, including the sober Aristotle, that all finite beings wish and strive, so far as their natures allow, to be absolutely perfect, or divine.

This is not the place to consider what might be said for or against such a view. It should be noted, however, that though Kant sees the idea of complete freedom from the inclinations as an archetype of practical reason (the idea of holiness) (KpV 32-33, 83), he clearly states that such freedom is not possible for a created and therefore dependent being (KpV 84; cf. 118) and moreover warns against the fanatical presumption (which he attributes to the Stoics, among others) that it can be achieved, maintaining that it should instead be recognized by such a being as a "fixed yet unattainable goal of its striving" (KpV 84-86). Kant's insistence that such a presumption is vain and dangerous does not, of course, by itself imply that he sees anything positive in the inclinations. But appreciating this insistence does make it easier to notice that his attribution to every rational being of a wish to be completely free from the inclinations does not preclude a view of the inclinations that accords to them, or to some of them, a worth that is tied to the place they occupy in the life-economy of a dependent being. That such a view is not ruled out is also apparent from the fact that Kant denies merely that the inclinations have a worth that is *absolute*, or *unconditioned*: to claim, as he does, that the inclinations are very far from having an absolute worth making them worthy of being wished for is not to claim that none of them have any worth in relation to the existence of a dependent being.

Clear indications that Kant does in fact hold that inclinations can have a positive worth are provided by certain statements in his later writings that have recently been pointed out by some of his more sympathetic

interpreters (e.g., Baron 1995, pp. 199-200; Wood 1999, p. 123). One of them is particularly striking: "Natural inclinations are *considered in themselves good*, that is, not to be rejected, and it is not only in vain, but it would also be harmful and blameworthy to want [*wollen*] to extirpate them" (R 58). Here again we have a warning to moral enthusiasts who would advocate the mortification of sensible desire, but this time it is accompanied by an explicit attribution of worth to certain inclinations, namely, *natural* inclinations, or inclinations based on true natural needs of human life. This worth, however, is clearly not absolute, nor is it equivalent to goodness all things considered. So what is Kant claiming in saying natural inclinations are good "considered in themselves"? Here it will help to draw on his idea of a natural end and the associated notion of a natural need. A natural end – a plant, for example, or an animal – is a self-organizing product of nature, which produces itself in the species, in the individual as a whole, and in each of its mutually furthering parts (KU §§64-66). Because a natural end is a dependent being, however, its self-furtherance is contingent on the satisfaction of certain natural needs, which, though not satisfiable in all possible conditions, are necessarily compatible, given that the end is possible. Since the human being is a natural end, it follows that its natural needs are consistent, that our species, as Kant says, is "in thorough agreement with itself" in respect of them (KU 430). Accordingly, the natural inclinations, being based on these needs, are likewise consistent as well as necessary, and so to that extent they qualify as subjective material conditions of the highest good – specifically of the happiness necessarily belonging to it – and hence as objects of practical knowledge, suitable for practical attention and cultivation (cf. KU 432). Thus, the natural inclinations are practically cognizable in their original, essential relation to human life in general, prior to any consideration of them in the state into which they may develop or misdevelop in a particular case through the choices made by the individual person to whom they belong, and they are accordingly good considered in themselves.[30] This practical knowledge can in turn support further practical judgments in which other things are deemed good as means; as Kant observes, human skill, industry, and their fruits acquire their "market price" through their practically cognized relation to these "universal human inclinations and needs" (G 434-435).

[30] Similarly, the gifts of nature and of fortune that Kant contrasts with the good will at the beginning of the *Groundwork* are all good considered in themselves (as Kant implies in calling them "gifts"). Whether in a particular case such a gift is good all things considered depends on how it is used by the will.

2. Elsewhere in the *Groundwork* and in the second *Critique* passages can be found that provide further context for Kant's claim that all rational beings wish to be free from the inclinations. These passages suggest that this claim can also be read on a second level, as employing a weaker sense of 'free', in which complete freedom from the inclinations is indeed attainable, at least in principle, notwithstanding the impossibility of their extirpation. Read on this level, Kant's claim expresses essentially the same view as the one we find articulated in this similar passage from Section III of the *Groundwork*:

> There is no one – not even the most hardened scoundrel, if only he is otherwise accustomed to use reason – who, when one sets before him examples of sincerity of purpose, of steadfastness in observing good maxims, of sympathy and general benevolence (even combined with great sacrifices of advantages and comfort), does not wish [*wünsche*] that he might also be so disposed. He cannot indeed bring this about in himself, though only because of his inclinations and impulses; yet at the same time he wishes [*wünscht*] to be free [*frei*] from such inclinations which are burdensome even to him. (G 454)

Here again, as in the earlier passage, Kant speaks of the wish to be free from the inclinations and proclaims it to be shared by everyone – even the most hardened scoundrel. But in this passage it is clear that the sort of freedom from inclinations that Kant says everyone wishes for is not one whose achievement would require the extirpation of sensible desire, but instead a type of freedom that is distinctively characteristic of the virtuous person. Such freedom, as Kant later describes it, is an independence of one's power of choice from all *influence* from feelings and inclinations, that is, from the efficacy an inclination can have in respect of the power of choice to the extent that it is strong enough to interfere with reason's determination of that power, either by leading to choices contrary to the moral law or by making choice in accordance with it difficult.[31] To achieve complete freedom from the inclinations in this sense would be to achieve complete virtue, in which we "release" ourselves "from the impetuous importunity of the inclinations to such an

[31] Cf. KpV 118, where it is said that in this freedom "one's determination of the will can hold itself free from the influence of inclinations and needs." There is yet another very similar passage in the second *Critique* (KpV 155-161) in which it is made explicit that the freedom in question in the passage at G 454 is *inner* freedom (the wish is mentioned at KpV 156 and the freedom is characterized as inner at KpV 161); and from related passages in the *Metaphysics of Morals* it can be seen, more specifically, that Kant has in mind inner freedom in a *developed* form, as a kind of *strength*. The sort of freedom Kant has in view in these passages is discussed in Engstrom (2002).

extent that none of them, not even the dearest, has influence on a resolution for which we are now to make use of our reason" (KpV 161).

Running through all these passages, then, we find a view of the inclinations that might be summarized as follows: The *natural* inclinations are good considered in themselves, since their natural function of satisfying the natural needs on which they are based enables them to be practically cognized in relation to the highest good. But no inclination is good absolutely. For on account of their self-strengthening efficacy as sensible desires the inclinations are "sources of needs" – i.e., able to grow so strong as to press "importunately" on the power of choice, thus generating new, acquired needs – and hence are liable to influence that power, interfering with its exercise according to reason, whereby they come to be burdensome to a rational being. Moreover, for much the same reason they become especially burdensome even on their own terms in a life devoted to the pursuit of their satisfaction. For in such a life inclinations tend to arise that lack the moderating relation to instinct characteristic of the natural inclinations, and (as Diogenes, Plato, and Rousseau also maintained) when indulged these inclinations not only grow stronger but also (owing to the internal connection between gratification and pain) hold out ever larger cups (and hence begin fighting among themselves) while also leading to boredom and a resulting restless succession of importunings, first by one inclination and then by another, and so become increasingly difficult to satisfy; hence the inclinations themselves "wrest" from a rational being the wish to be relieved of them as greater maturity and experienced reflection reveal the aim of such a life – bare happiness – to be inherently unattainable (KpV 118; cf. G 395-396, KU 430). In a virtuous person, however, where freedom from the inclinations has been achieved, the natural inclinations, though of course still present, are no longer the source of needs and so no longer burdensome; in such a person they are not only good considered in themselves, but good all things considered. At bottom, then, "the universal wish of every rational being" to be "completely free" from the inclinations is nothing but the wish to be virtuous, present as the good will in every rational being under the moral law, which in its most primitive form is a wish for the fixed yet unattainable ideal of complete rational perfection.

University of Pittsburgh
Department of Philosophy
Pittsburgh, PA 15260, USA
e-mail: engstrom@pitt.edu

REFERENCES

Baron, M. (1995). *Kantian Ethics Almost Without Apology*. Ithaca, NY: Cornell University Press.

Beck, L.W. (1960). *A Commentary on Kant's* Critique of Practical Reason. Chicago: University of Chicago Press.

Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell Publishers.

Engstrom, S. (2002). The Inner Freedom of Virtue. In: M. Timmons (ed.), *Kant's* Metaphysics of Morals*: Interpretative Essays*, pp. 289-315. Oxford: Oxford University Press.

Foot, P. (1978). *Virtues and Vices*. Berkeley: University of California Press.

Herman, B. (1993). *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.

Herman, B. (2001). Rethinking Kant's Hedonism. In: A. Byrne, R. Stalnaker and R. Wedgwood (eds.), *Fact and Value: Essays on Ethics and Metaphysics for Judith Jarvis Thomson*, pp. 129-153. Cambridge: MIT Press.

Hume, D. (1739). *A Treatise of Human Nature*.

Irwin, T.H. (1984). Morality and Personality: Kant and Green. In: A. Wood (ed.), *Self and Nature in Kant's Philosophy*, pp. 31-56. Ithaca, NY: Cornell University Press.

Kant, I. (1902-). *Kants gesammelte Schriften, herausgegeben von der Deutschen* [formerly *Königlich Preußischen*] *Akademie der Wissenschaften*. 29 vols. Berlin: De Gruyter (and predecessors).

Korsgaard, C.M. (1997). From Duty and for the Sake of the Noble: Aristotle and Kant on Morally Good Action. In: S. Engstrom and J. Whiting (eds.), *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty*, pp. 203-236. Cambridge: Cambridge University Press.

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

Sherman, N. (1997). *Making a Necessity of Virtue*. Cambridge: Cambridge University Press.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell Publishers.

Wood, A. (1999). *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Steven Arkonovich

GOALS, WISHES, AND REASONS FOR ACTION

1. Hybrid Humeanism

Suppose I want to see all 12 hours of the marathon PBS presentation "The Three Tenors: Again and Again and Again." I don't want to miss a single minute, but I don't know when it airs. So I buy TV Guide. We can explain my purchase by citing my desire to see the special, together with the belief that buying TV Guide will aid me in that end.

This example contains elements of a very familiar view about the proper explanation and justification of intentional action. Let represent some verb of action. Then: if an agent 's he must have wanted to ; and his desire to plays an essential role in the correct intentional explanation of his -ing. Further, the desire to (or one relevantly related to -ing) will have an essential role to play in *justifying* his of -ing; He has a good reason to just because he wanted to. Let us call this conjunction of claims Humeanism. Humeanism, then, consists in these three claims:

(1) If *A* 's intentionally, then *A* desires to .
(2) If *A* 's intentionally, *A*'s desire to motivates and so explains *A*'s -ing.
(3) If *A* has a reason to , then *A* has a desire to (or a relevantly related one) which justifies *A*'s -ing.

Defenders of Humeanism will suppose, rightly I think, that (1) is undeniable. Even in those cases in which the action undertaken is quite unpleasant, we cannot, if we are to see the action as intentional, afford to deny (1). Suppose Jones goes out into the bitter cold to keep an appointment with his dentist. He goes out even though he could stay home and avoid the unpleasant trip and the resulting pain. Surely it is correct to say in some sense – a sense adequate to secure (1) – that Jones wants to go out.

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 161-184. Amsterdam/New York, NY: Rodopi, 2007.

(2) expresses the Humean theory of *motivating* reasons. Motivating reasons are appealed to in the explanation of action; they are what move an agent. Someone can have a motivating reason that does not move her all the way to action. But motivating reasons are, by their nature, potentially explanatory of the agent's actions. According to the Humean theory of motivating reasons, desire is what moves people to action. Therefore all motivating reasons essentially involve desires. We should notice that (2) is stronger than (1): (1) follows from (2). For (1) says that someone who acts intentionally must *have* a desire to so act. This could be true even if what *moves* the agent is, for example, a belief. By contrast, (2) says that the very state that motivates all intentional action is, or at least partly is, a desire.

Let us say that (3) expresses the Humean theory of *normative* reasons. Normative reasons are the reasons we refer to in evaluating an agent's action from some normative standpoint. They are the reasons we appeal to when we say that an agent has good or bad, sufficient or insufficient, reasons for acting. An agent can have normative reasons on which he does not act: he might not know about them, or he might know about them but decide to act on other, better reasons. And even where an agent is fully aware of which reasons are his best reasons, he may not be moved by them. He may be in such a state (of depression, or elation, for example) that his awareness has no effect. The proper understanding of how this can occur is a central theme of the present paper.

Those who want to reject Humeanism have several options, but the most popular are the rejection of (3) alone, or the rejection of both (2) and (3).[1] One might want to reject (3) for a couple of reasons. Minimally, one might want to claim that (3) restricts the range of reasons by tying them necessarily to things we desire. So, while one might admit that desires *sometimes* provide good reasons to act, one could at the same time insist that there are other sources of justification apart from what a particular agent happens to want: the needs of others, or the claims of justice, for example. In the same vein, though more radically, one may insist that desires are not even the right sort of thing ever to provide

[1] Few will want to bother about (1), because even in accepting it one is not committed to saying anything about the explanatory or justificatory power of desires. Indeed, all one seems committed to in accepting (1) is the linguistic propriety of saying that an agent who acts intentionally must have wanted to so act. Desires so ascribed have been called merely "formal" desires, or "desires in the philosopher's sense." To ascribe a merely formal desire is simply to say that, indeed, the action under questions was intentional.

reasons for action. On this view, tying reasons to desires isn't just too restrictive, it is utterly misguided.

Warren Quinn has advanced this more radical rejection of Humeanism. Quinn asks us to imagine someone who is disposed to turn on radios, but who does not have any tendency to value these actions. He neither thinks that there is good to be achieved in having radios on, nor in his act of turning them on. He is simply disposed to do so. On many contemporary views of desires this suffices to attribute a desire to the person exhibiting this odd behavior. For on these views desires are essentially (and merely) dispositional states. Quinn asks,

> how can a non-cognitive functional state whose central significance in this context is to help explain our tendency to act toward a certain end, or in accordance with a certain principle, *rationalize* our pursuit of the end or our deference to the principle. How can the fact that we are set up to go in a certain direction make it (even *prima facie*) rational to decide to go in that direction? How can it even contribute to its rationality? (Quinn 1994, p. 236)

Quinn's answer is that it cannot, at least as long as desires are conceived as mere functional states, not essentially tied to evaluation.

One might want to pursue a more thoroughgoing rejection of Humeanism by denying (2) – the Humean theory of motivating reasons – along with (3). Indeed, this more radical rejection of Humeanism has recently been gaining favor (Scanlon 1998, Korsgaard 1996). But such views, it seems to me, are difficult to accept precisely because they fail to give a convincing desire-independent account of motivating reasons. If that is right – and I will not argue for it here – then the best hope for resisting Humeanism will be on more moderate grounds, in the rejection of (3) alone. The aim of this paper is to argue against this middle position, thereby forcing the choice between full-blown Humeanism or a thoroughgoing rejection of the view. As I said, if one finds this thoroughgoing rejection implausible, then Humeanism is the view to adopt at the end of the day.

The moderate view, then, would be one which remains loyal to (1) and (2) while jettisoning (3). On its face, this seems like an attractive option. For if desires seem to be poor candidates to be "justifiers," there is every reason to think that are very good candidates to be "motivators." Supporters of the moderate view could accept that desires are what motivate action and while denying that desires play a necessary role in the justification of action.

Michael Smith has offered a very influential defense of this moderate rejection of Humeanism (Smith 1995, 1992). He argues that we must

accept the Humean conception of motivating reasons while accepting a non-Humean account of normative reasons. I will call this moderate view "hybrid Humeanism," and I will argue that such a view cannot succeed. I also show that Smith has not provided any *incentive* to construct a hybrid Humeanism. For Smith thinks that we need a compromise view to if we are to explain common failures of rational motivation. Against this, I suggest that a more sophisticated understanding of desire not only allows Humeanism to give an account of these failures, it allows Humeanism to give a better account than the one Smith offers.

2. The Teleological Argument

I said that many people find the Humean theory of motivating reasons very plausible. The strongest argument for the theory is Smith's own (Smith 1987, 1995). His argument starts with the idea that reason explanations are essentially teleological. Teleological explanations, it is said, make an agent's action intelligible by showing the agent as pursuing a goal. We make my buying of the TV Guide understandable when we indicate that my aim is to know when the PBS special airs. According to Smith, this teleological character can be preserved only on the Humean assumption that the explanation of an action makes reference to an agent's desires.

The alternative explanatory scheme that Smith opposes allows that at least sometimes an agent's beliefs about some aspect of his action (usually its desirability, or goodness, or that it would help a friend, etc.) might be the sole motivating factor in moving the agent to act. On this alternate, "belief-based" account of motivation it may suffice to explain my action (say, my going to the dentist) that I believe it will lead to my greater good. What is not needed, according to these accounts, is some appeal to a desire of mine (say, a desire to do things that will lead to my greater good).

Central to understanding Smith's opposition to these "belief-based" motivational explanations is the notion of "direction-of-fit." Mental states are said to have one (and only one) of two directions of fit:

(i) mind-to-world,
(ii) world-to-mind.

Mental states of type (i) are such that their propositional content is supposed to "fit" the world. Mental states of type (ii), on the other hand, are such that the world is to be made to "fit" their content.

Beliefs are the paradigm of type (i) mental states. Beliefs aim at the true, and so it is appropriate for someone's beliefs to be "acquired" and "discarded" according to whether what is believed is, or is not, the case. Desires are the paradigmatic type (ii) mental states. Desires aim not at truth, but at satisfaction. They differ from beliefs in that it is appropriate for the content of desires not to match the world. It is the world that should change so as to fit the content of the desire.[2] With this distinction in the background, we can set out Smith's short, but very influential, argument:

(4) Having a motivating reason *is, inter alia*, having a goal.
(5) Having a goal *is* being in a state with which the world must fit.
(6) Being in a state with which the world must fit *is* desiring.[3]

The argument seeks to make clear the conceptual connections between desires and motivation via the ideas of having a goal and direction of fit. We know simply in virtue of what it is to be motivated that it is (in part) to have a desire.

One salient feature of this argument is that it gives rather special senses to some of the terms it uses. This is evident, I think, with respect to the term 'desire', and even more evident with respect to the phrase 'having a goal'. According to the teleological argument, it suffices to have a desire that the agent have (or that she be in) a state with a certain direction of fit. Smith is anxious to affirm that this allows a number of mental states that we might normally distinguish from desire to fall under the rubric of desire. For instance, hopes and wishes will count as desires according to the teleological argument, even though we might want to distinguish them in everyday speech. Smith imagines someone objecting to his argument precisely on the grounds that desires are not wishes or hopes. He replies,

> if desire is not a suitably broad category of mental state to encompass all of those states with the appropriate direction of fit, then the Humean may

---

[2] In discussing "direction of fit" I am being deliberately unclear about whether the "appropriateness" or "inappropriateness" attaches to the behavior of the person who has the beliefs and desires, or whether it attaches to the mental states themselves, e.g., whether it somehow goes against their nature. There are serious difficulties in understanding the idea of "direction of fit," and not only with respect to this ambiguity. Richard Wollheim provides a sensitive and sensible discussion of the idea of "direction of fit" in Wollheim (1999). See also, Schueler (1991) and Humberstone (1992).

[3] Smith (1987, p. 55). Also, Smith (1995, p. 116). The copulas are italicized in the former and dropped in the later.

> simply define the term "pro-attitude" to mean "psychological state with which the world must fit." (Smith 1995, p. 117)

The argument also gives a rather special sense to the phrase 'having a goal'. Let us say that somebody wishes for something, and that what he wishes for is impossible: he wishes for a different past. Such a wish obviously has a very tenuous connection to anything we would normally call the agent's "goals." This is so even when the wish is for an action on the part of the agent: he wishes that he had swerved to avoid a dog in the road. It is hard to imagine a case where we would say of such a person that "he (now) has having swerved as a goal." Notice, though, that as the phrase 'having a goal' occurs in the teleological argument we must say that the agent has this as a goal. This is because wishing is a form of desiring, and so wishing is a state with which the world must fit. And by premise (4), having (or being in) a state with which the world must fit suffices to have a goal.

That it leads to these odd verbal consequences is not in itself an objection to the teleological argument. In any case, I have not introduced the argument because it is undeniably sound, but because it embodies a fairly clear articulation of the intuitions behind the Humean theory of motivation. It expresses a very common way of conceptualizing desires, and especially the distinction between desires and beliefs. I highlight the odd verbal implications of the argument now because such notions as "having a goal" will recur, and it will be important to keep in mind their rather technical sense.

3. The First Argument for Hybrid Humeanism

Let us turn to the heart of the compromise view developed and defended by Michael Smith. We have just seen the grounds on which Smith (and many others) accepts a desire-based theory of motivating reasons. What is distinctive about a compromise view, however, is that it also endorses a desire-independent theory of normative reasons. Smith begins his defense of this aspect of his view by offering a challenge to *any* theory of normative reasons. He thinks that the challenge can be met only by a belief-based theory of normative reasons.

When we deliberate and reach conclusions about what we have reason to do, we are often motivated to act accordingly. Less often, we are not so motivated. Smith's challenge is to explain "how it can be that accepting normative reason claims can both be *bound up with* having desires and yet *come apart from* having desires. In other words, the

problem is to explain how deliberation on the basis of our values can be practical in its issue *to just the extent that it is*" (Smith 1995, p. 136, original emphasis).[4] A successful theory of normative reasons will have to take the right measure of the gap that exists between our (acceptance of) normative reasons and our motivations.

Consider desire-based accounts first. These Humean accounts will have a hard time meeting his challenge. For if we think that valuing is simply a matter of desiring, then although it is clear how our values can motivate us, we no longer have any explanation of how they could *fail* to motivate us. If valuing just is desiring, then an agent who values some course of action just will be motivated, to some degree, to pursue it since desires are intrinsically motivational states. But that we can fail to be motivated cannot be doubted. Smith thinks that we should be convinced by the following remarks of Michael Stocker:

> Through spiritual of physical tiredness, though *accidie*, through weakness of body, through illness, through general apathy, through despair, through inability to concentrate, through a feeling of uselessness or futility, and so on, one may feel less and less motivated to seek what is good. One's lessened desire need not signal, much less be the product of, the fact that, or one's belief that, there is less good to be obtained or produced, as in the case of a universal Weltschmerz. Indeed, a frequent added defect of being in such "depressions" is that one sees all the good to be won or saved and one lacks the will, interest, desire or strength. (Stocker 1979, p. 774)

Smith writes as if the examples simply clinch the case against desire-based accounts, and in that he is surely too hasty. For he seems to assume that these examples show that an agent can value doing something in the absence of *any* desire to do it. But surely some defenders of desire-based accounts will say that this begs the question. Perhaps all that is needed to account for the maladies that Stocker points to is the supposition that the agents are less motivated to pursue what they value, or have strong countervailing motivations. It is not even clear that Stocker understands the examples in the way Smith supposes (Stocker does write about *lessened* desire, after all). So it is certainly possible to read these examples in a way that does not rule out desire-based accounts.

It surely is possible, but I believe that Smith is correct to say that the burden is on desire-based accounts to present a plausible and natural

---

[4] As is apparent in this quote, Smith does not distinguish between valuing a course of action and believing that one has reason to pursue it. This distinction will be discussed in some detail.

reading of these examples. For what the examples present us with, on their face, are people whose appropriate motivation to pursue what they value is absent, and not merely reduced or overwhelmed by the force of contrary desires. So there is a genuine challenge for desire-based accounts here. We will return to the question of how those accounts might meet it later.

Let us turn, then, to belief-based accounts. If we think that valuing a course of action is a matter of believing it valuable, then we have an easy explanation of why an agent can be unmoved to pursue what he values. Since desires and beliefs are, as Hume says, "distinct existences," there is no problem in having one without the other. The difficulty for these belief-based accounts is to explain how deliberation on the basis of our values could *ever* motivate. Since the teleological argument gives us reason to think beliefs can never motivate, it seems we could have no explanation of how simply believing something valuable can affect what we do. Whoever understands valuing to be a matter of believing may be able to explain a gap between our reasons and our motives, but she seems to have no way to close the gap.

Smith thinks that there really is no way to open the gap on desire-based accounts of valuing, so instead he tries to find a way for belief-based accounts to close it. He thinks this could be done if there were a norm of rationality linking evaluative beliefs with desires. A norm like this would say what a rational agent desires given that he has certain evaluative beliefs. Thus Smith's project is to construct an account of normative reasons centered around the idea of a "rationalized desire," or a desire had for a reason.[5] What we need, according to Smith, is a norm like this:

(C) If an agent accepts that he has a normative reason to , he rationally should desire to (Smith 1995, p. 148).

If (C) really is a norm of rationality, then we could allow that valuing is a matter of believing valuable, and so see why values and desires can come apart (as we should be able to if they are "distinct existences"). But at the same time we will have an explanation of how an agent's values can make a difference in what she does. An agent's desires will conform

[5] Thomas Nagel's account of normative reasons, as presented in Nagel (1970) is also worked out around the idea of a "rationalized desire." Nagel's view is not a compromise view, however, since he rejects (2). Furthermore, Nagel's notion of a rationalized desire is importantly different from Smith's. For one thing, Nagel's rationalized desires cannot explain actions, whereas this is precisely what Smith needs them to do.

to her "reasons beliefs" just to the extent that she is rational. Therefore (C) represents a *defeasible* connection between an agent's belief about her reasons and her desires.

We must be careful here. Proponents of desire-based theories of normative reasons need not deny every interpretation of (C); they can accept that a rational agent desires to do what he believes he has normative reason to do. However, they will explain the truth of (C) by saying that what an agent has reason to do is just a function of what that agent already desires. Now Smith insists that (C) should be interpreted to represent a defeasible connection between reasons and desires, which the proposed Humean understanding of (C) doesn't allow. But it seems that the Humean can accept that (C) draws a defeasible connection. For, strictly speaking, (C) expresses a relation between the *acceptance* of certain statements and the having of certain desires. And even on desire-based view an agent could *accept* claims about what he values (or has normative reason to do) without having the appropriate desires, since he could be mistaken about what he values.[6]

The real challenge for Smith in distinguishing his compromise view from desire-based theories is to come up with a desire-independent understanding of (C). This will involve defending a desire-independent account of the conditions under which an agent has a normative reason. If he can do this we will we have an explanation of how valuing a course of action makes just the difference that it does in the agent's desires. It will make just the difference that it does not because valuing is desiring, but because valuing causes and rationalizes the desires of a rational agent.

As I read Smith, he rests his defense of his compromise views on certain "platitudes" about the relations among values and reasons and

---

[6] Smith is not always careful about this. This is particularly clear in his discussion of the distinction between Humeans and anti-Humeans about motivating reasons. According to Smith, Humeans about motivating reasons, like himself, think that beliefs about what is right can motivate only with the help of desires. Anti-Humeans (like John McDowell, Thomas Nagel and Tim Scanlon) think that beliefs about what is right can motivate by themselves, unaided by desire. Thus, he says, Humeans about motivating reasons can allow that agents who believe that -ing is right may not desire to . But Anti-Humeans must accept that "it is impossible for agents who are in a belief-like state to the effect that their -ing is right not to be in a desire-like state to the effect that they ." This may or may not be a good inference. But he is surely wrong to say that in deciding the matter "[e]verything turns on how we should interpret the idea that moral *judgment* is essentially practical. Is the idea that someone who *judges* it right to act in a certain way is motivated to act accordingly *simpliciter*? Or is the idea rather that someone who *judges* it right to act in a certain way is motivated to act accordingly *ceterus paribus*?" See Smith (1995).

desires. Smith offers what he considers the following two platitudinous biconditionals:

(P1) *A* accepts that he has a normative reason to iff *A* accepts that his -ing is valuable (Smith 1992, p. 329).[7]

(P2) *A* has a normative reason to iff *A* would desire to if he were rational (Smith 1992, p. 329; and 1995, p. 150).

(P1) is necessary to Smith's argument since the original challenge was to explain the defeasible connection between our values and desires. If that explanation is going to go via claims about normative reasons, there must be a very close connection between our values and our reasons. (P1) draws that connection in the closest way possible. (P2) then connects our normative reasons to our rational desires. Importantly, Smith argues that we should analyze (P2) in terms of what he thinks are further platitudes about advice. Roughly, what it is rational for me to do is what I would do if I followed the best advice. And (again, platitudinously) the best advice would come from myself, suitably idealized. Normative reason claims like those in (C), therefore, should be read as claims about what one's ideally rational self would advise one's (non-idealized) self to do in one's present (non-idealized) circumstances. Thus the articulation and defense of (C) rests on the fact that it coheres with (P2) and the associated platitudes about advice. (C) represents a defeasible connection between normative reasons and desire since Smith supposes that what one's ideally rational self would advise one's actual self to do is not necessarily connected with what one actually wants. That is why it turns out that normative reason claims are desire independent, and can obviously come apart from one's desires. But to the extent that one is rational, one will desire to do what one's ideally rational self would advise one to do. That is why, to the extent that one is rational, one will desire in accordance with what one has normative reason to do. That is how a compromise view can meet Smith's challenge.

This argument is only as strong as the argument that (P1) and (P2) are platitudes. There is considerable room for doubt here. Reading from left

---

[7] (P1) does not turn up in Smith's book (1995) as an explicitly stated platitude, though it does in his earlier paper (1992). However, Smith still needs the platitude to justify his shift from talk about what an agent believes she values to talk about what an agent believes she has reason to do. Furthermore, it is clear that in *The Moral Problem* Smith assumes the two things to be equivalent. He says, for example, "if I think that it is desirable that I write a book – or, equivalently, if I think that I have normative reason to do so – and I think that I can write a book by typing these words, then we can redescribe my attitudes in these terms: I *value* writing a book [. . .]" (1995, p. 132).

to right, (P1) says that if I accept that I have a normative reason to then I find my -ing valuable. But could I not find myself in the unfortunate circumstance in which, although I have reason to choose between two courses of action, I do not think either of the options open to me are valuable? Perhaps I find myself in the position of having to choose between a mocking by Sue or an insulting by Sally. As I reflect on my choice, I realize that each option will be humiliating in its own (perhaps incommensurable) way. It would not be natural describe my deliberations when faced with this dilemma as concerning which outcome I found more valuable, my being mocked or my being insulted. Rather, we should say that I was trying to decide which was less awful. And "less awful" simply doesn't imply "more valuable."

It might still be insisted that there has to be something in the situation that I find valuable. How else can we give substance to the idea that I am deliberating, or choosing, as opposed to merely picking? I just suggested that a natural way to conceive of the situation is as trying to discover which action of mine would be less awful. But let us, for the sake of argument, admit that there must be something more going on in my deliberations to make this a case of choosing and not just of picking. Suppose I know how much Sue loves to mock. If I have nothing against Sue (though it is hard to see why I wouldn't), I might find her enjoyment in mocking me valuable. So I choose a mocking over an insulting. It does not follow that I value my acting so as to bring upon myself a mocking. What happens is that I find myself having to choose between these awful alternatives and I latch on to a relatively *ad hoc* consideration to help me decide. We do this all the time. Something like this goes on when, not knowing which can of green beans to buy, I decide on the basis of the label. I can do this without thinking that the label indicates that the contents of the can have the virtues appropriate to green beans.[8] In the same way, I might have noticed that Sue was wearing green while Sally was wearing blue. If I like green more than blue, I may let that fact guide my choice and put up with a mocking instead of an insulting. Again, I reach for some *ad hoc* consideration, and again we could not conclude from this that I value my being mocked; or value my acting so as to bring

[8] The claim is not that the *ad hoc* consideration provides me with a *justification* for my choice. What I have reason to do is choose *A* or *B* and the *ad hoc* consideration, though it helps me choose, does not provide additional reasons for the choice. The claim here is only that there is a way of choosing on the basis of a consideration that does not commit the agent to finding the choice valuable. I would like to thank Sergio Tenenbaum for pressing me to clarify this point.

it about that I am mocked; or value either of these things more than I value the insulting alternatives. So I don't think we have any reason to be convinced that thinking one has a reason to entails thinking that one's -ing is valuable.

Read in the other direction, the "platitude" says that if I believe my -ing is valuable, then I believe that I have normative reason to . But again this is not true. Perhaps what I value is my painting a fine painting. But given my late age or the progress of my disease, or simply my lack of talent, that fact does not directly give me reason to do much at all. (It certainly does not give me any reason to *try* to paint such a painting.) Perhaps we think that I cannot value my painting a fine painting precisely because it is not something I can pull off. There are two responses available. The first response is to deny the charge. I see no reason to think that one cannot value one's doing something that one cannot do, and so something one could have no reason to do. Suppose that I have every confidence that I can paint a very fine painting. And so, with this confidence in my abilities, I start out on this project. Could it not gradually dawn on me, as I labor over the canvas, that I am just not up to the task? My confidence was misplaced. Are we constrained to believe that as my doubts increase they must invariably be accompanied by a corresponding drop in my estimation of the worth of my completing the project I set out on, a project I now think is impossible? We are not. At most we should grant that it would be misleading for me to go about telling people that I value doing something if my so doing is (as I have come to think) impossible. Others might be lead to think that it *was* possible, or at least that I thought it was possible. But it would be incorrect, as it often is, to infer from this conversational convention something that is a substantive psychological conclusion, namely, what I, or anyone else, can find valuable.

The second response is to see that there is a distinction between valuing and having a reason even in those cases where impossibility is not at issue. My reason for valuing painting a fine painting may be that I believe that anyone's painting such a painting is valuable. After all, fine paintings are valuable in themselves, and so most anything that leads to their creation is valuable. Let us suppose that I have every reason to be confident that, with proper training I could produce a very good painting, indeed. Still, the smell of paint gives me headaches and makes me irritable. Anyway, I have no interest in undertaking such a daunting course of study; my real love is science. So I have no reason to pursue a course of action I believe is valuable.

In short, it seems not a logically necessary feature of any rational life, much less a point that can be established by considering the grammar of certain sentences, that one's values feed in a direct way into what one has reason to do. That there is a close connection between what one values and what one has reason to do is a condition of a minimally fortunate life, something quite beyond the power of logic or grammar to guarantee.

This points up a genuine weakness in Smith's argument. For it is no doubt true that on many occasions what I value does have consequences for what I have reason to do. But, as we have just seen, there is certainly reason to deny that I will accept that I have reason to pursue or promote everything that I admit it is valuable that I pursue or promote. What is required, then, is an account of the conditions under which what I accept as valuable will give me reason to act. A natural place to look for such an account is in the desires of the agent. So even if valuing is believing, we may still allow that accepting what one has normative reason to do depends upon one's desires. Desires could mediate between the claims of value and the claims of rationality.[9]

Smith might admit that he was too hasty in identifying valuing with having normative reasons. But he would not have to accept the suggestion just made, namely, that having a normative reason is necessarily connected with having an appropriate desire. He could say: "It is true that we cannot assimilate normative reasons and values. Still, it is undeniable that sometimes we do deliberate on the basis of our values and reach conclusions about what we have reason to do. And even in these cases, agents can be afflicted with the sorts of ills that Stocker describes. So we can accept that we have normative reasons without being motivated to act, and that shows that valuing must be a matter of believing. We just need to find the right content for the belief, a content that will justify (C) and show why it represents a defeasible connection between accepting claims about one's reasons and desiring accordingly."

So let us now look at (P2). Again, Smith wants us to read (P2) as concerning a desire "[. . .] I would have, if I were rational, about what I am to do in my actual circumstances" (Smith 1992, p. 347). We are to imagine our idealized, completely rational self, and then answer the question, What would he desire that I do in my actual (not completely rational) circumstances?

This is not the place for a detailed examination of Smith's "ideal advisor" account. But we can note some general concerns we should have

---

[9] Indeed, this is the approach that Bernard Williams seems sympathetic to. See Williams (1995).

about this "platitude." For instance, there are well known problems with the kind of counterfactual claims in (P2). Who knows what I, or someone as much like me as possible, would desire if he (or I) were fully informed and cognitively perfect? Perhaps that sort of cognitive achievement would lead to utter boredom, so that he simply wouldn't care what I did. Relatedly, what is the connection between what my ideally rational self would *advise* me to do (or to want) and what he himself would *desire* that I do or want? Why do we get to suppose that my ideal self's desires (about non-idealized me) would be expressed directly in some bit of advice? On Smith's account we are really supposed to take the motives and actions of our idealized counterpart seriously – claims about what they would be is what is delivering answers about our actual reasons. But these claims, it seems, are largely undefended and so it is difficult to evaluate whether the utterances or desires of my ideal self can form part of a plausible account of my reasons for action.

Furthermore, there is reason to be concerned about the justification Smith provides for (P2). Recall the Smith thinks that (P2) is itself supposed to follow from such platitudes.

> As I see it, [(P2)] is related to a whole host of platitudes about *advice*. If you are unsure about what to do in some situation, how should you go about deciding what to do? The answer is that you should tap into the wisdom of the folk; you should ask for advice. But you shouldn't just ask any old person for advice. You should ask someone better suited than yourself to know what you should do, someone who knows you well. (Smith 1995, p. 151)

My suggestion is that Smith's account is not at all an articulation or extension of various platitudinous connections we are willing to draw between reasons and advice, and so it gains no support thereby.

What is a commonplace – and this is the situation with which Smith begins – is that sometimes when I am unsure about what to do I should inform myself. One good way to do this is to seek advice from someone who, as Smith puts it, is "better situated" than I am. It is easy to imagine an example where the connection between my reasons and good advice is straightforward. If I am hunting lions on the savanna, I will have a keen interest in knowing where lions might be lurking. In this situation, I might want to get the advice of the friendly man sitting in the tall tree nearby. He is, in an uncontroversial sense, better situated than I am to gather the information that I seek. He certainly could be a very good advisor indeed, and we can see why my reasons are tied to what he would tell me to do.

But we travel a considerable distance when we leave behind such an example – with its unexceptionable understanding of a "better situated" advisor – and advance the idea that the best situated advisor will always be me, transformed into a fully rational agent. This idea diverges from our uncontroversial example in several important ways. One difference, one Smith suggests, is that my idealized self will almost inevitably have a radically different set of concerns than I actually have. In the standard case (as in our example), an agent seeking advice wants to know what to do *given* her preferences. This means that she wants to know what to do given at least some of (and perhaps quite a lot of) what she already takes to count in favor of and against certain courses of action. Our agent will have various desires that point in different directions, and importantly, because the possible courses of action (as well as what counts in favor of them) are to a great extent settled, her standard deliberative question will be, "Do I desire *this*?" or "Should I do *that*?". Of course, there is always some room for the suggestion that the agent should reorder her preferences, or acquire or discard a particular preference. And less often there will be a great deal of room for such modifications. When there is, this will be reflected in the fact the natural question in deliberation will be "What sorts of desires should I have?" But it is doubtful that this question is, in general, the correct starting point for practical deliberation. Indeed, it has been noted that this question is most often appropriate when a person is in a state of deep conflict and confusion (Wollheim 1984, p. 167).

These worries about (P2) and (P1) indicate that Smith's argument for (C) does not receive strong support by cohering with these "platitudes." It is not just that the platitudes are not obviously true, and so are not what we might normally call platitudes. Rather, it is that they seem to draw dubious connections between our values and our normative reasons. Indeed, we might venture to say that what really is a platitude is that the relation between one's values and one's reasons is *complicated*. That platitude would lead us to think that any highly general biconditional relation between our values and reasons is likely to be false.

4. The Second Argument for Hybrid Humeanism

There is another line of argument that Smith might offer in defense of (C). Instead of offering an independent argument which, if successful, would provide us with a particular understanding of a norm that could meet his challenge, he might argue that there *must* be some such

understanding precisely because it is the only way to explain the gap. That is, he might say that the case in favor of (C), or some understanding of (C), gains significant support just because it could explain the defeasible connection that exists between our reasons (or values) and our motives. Indeed, this argument would, if successful, bolster the support for compromise views in general. After all, we should all admit that the "gap" to which Smith points us does need an explanation. If a belief-based theory can explain it, and a desire-based theory cannot, then that gives us strong reason to prefer the one theory to the other.

This new argument will be only as strong as the desire-based view is weak. But I do not think that Smith can show belief based-accounts of normative reasons to have any advantage over their desire-based counterparts in meeting Smith's challenge. Indeed, I will argue that even if we grant Smith's understanding of (C), his compromise view cannot provide us with the required explanation.

To see this we need only notice that, at least sometimes, what we have reason to do is based uncontroversially in our desires. Sometimes we have reason to act just because it gets us what we want. But even in these cases Stocker's ailments can prevent one from being motivated to do what one thinks one has reason to do. Such maladies can undermine our motivation to act on our reasons, even those reasons that are essentially connected to our desires. Assume that my moral reason to help the poor is a desire-independent reason, but that my reason to keep up my stamp collection is a desire-dependent reason. There is no plausibility in the assumption that when depression strikes, I must inevitably lose only my motivation to help the poor, while I must remain at least partly motivated to collect stamps. Nor is it plausible to think that I will necessarily stop believing that I have reason to continue either of those things. Smith's account has no way to accommodate these familiar facts about motivation.

It follows that Smith's account of the *irrationality* involved in these cases cannot be right. When my reasons are rooted in my desires, and I fail to be motivated appropriately, I am irrational. But what that irrationality cannot consist in is my failing to desire to do what I think I have reason to do since, *ex hypothesi*, the only reason I have to do it is that I want to. It is equally clear that my irrationality does not consist in lacking some further desire, and so lacking some further goal, e.g., the goal of pursuing my goals – or what is equivalent in this case, the goal of doing what I have reason to do. If I acquired that further goal, so that now I did have as a goal the pursuit of my goals, I would be no less irrational for that. My irrationality lies simply in not being appropriately

motivated to pursue my goals, since that is what I think I have reason to do.

These remarks show that Smith's account must fail when we turn our attention away from desire-based reasons and look at purported desire-independent reasons of the sort Smith is actually trying to defend. If adding to my psychology an additional desire, or "goal-directed" state, did not alleviate my irrationality when we were assuming that my reasons were desire-based, then adding such a state will not help when my reasons are desire-independent. However I arrive at the belief that I have reason to , it seems we can add as many of these goal-directed states as we would like. Still, if I am not motivated to pursue these goals, I remain irrational. So even if we grant Smith his purported norm of rationality he cannot meet his own challenge. Since the argument for the compromise view was based on its ability to meet the challenge, the argument fails. It has not explained "how deliberation on the basis of our values can be practical in its issue *to just the extent that it is*." That was the challenge Smith offered, and his account is not up to the challenge.[10]

The upshot is, I am suggesting, that there must be an important difference between having a goal and being motivated to pursue it. For any attempt to explain the "gap" that undeniably can open between our desire-based reasons and our motivation with respect to them will certainly need such a distinction. But once someone's desire-based reasons and motivations have come apart it is clear that we cannot lessen his irrationality by adding to his mental economy a mental state that embodies the having of a goal. He already has one of those, he is just not, as it were, doing enough about it.

Furthermore, the distinction between having a goal and being motivated to pursue it gives us grounds for doubting the teleological argument. For in light of that distinction, we can see that the teleological argument elides two corresponding notions of "goal-directed" state. One notion naturally characterizes particular mental states, i.e., desires. In this sense goal-directed states are to be contrasted with "truth-directed" states, i.e., beliefs. But, the objection continues, there is another sense of "goal-directed state" at work in the argument, which more naturally characterizes the general condition of the organism, as when we say an organism is, for example, seeking light or seeking truth. Obviously some organisms can be in goal-directed states in that sense – so that teleological explanation is appropriate – without being so in virtue of

[10] This argument of the last three paragraphs is presented in Arkonovich (2001).

having particular goal-directed mental states. And conversely, some organisms can have a goal-directed mental state without, as it were, being in a goal-directed state: people can have goals without being motivated to pursue them.

Smith wants to say, what is surely plausible, that in the case of the intentional pursuits of human agents, having a goal in the general sense is realized in virtue of having relevant, particular, goal-directed mental states. But the current objection contends that this could be denied. At least the teleological argument has not demonstrated it to be the case, for it does not even mark the distinction between the two ways we can understand the idea of a "goal-directed" state. I do not want to take a stand on whether we should reject the teleological argument altogether. It could well give an adequate account of the difference between desire and belief without giving an adequate account of desire. And it could still bring together in an intuitive way a set of considerations which support the Humean theory of motivation. We must conclude, however, that the conception of desire that it expresses is quite impoverished. The solution to Smith's challenge lies in finding the right ways in enrich it.

In the next section I will consider one way in which we must enrich our conception of desire: we must consider not just the object of the desire (what the desire is for), but also the role the desire plays within the psychology of the agent (what, as it were, the desire does for the agent). I will further suggest that in paying attention to the roles of desire we provide the Humean the tools with which to answer Smith's challenge. Second, to the extent that the conception of desire embedded in the teleological argument is pervasive among philosophers, a further implication of my argument will be that philosopher's must change their thinking in this arena. I believe that we cannot arrive at a satisfactory moral psychology if we rest content thinking of desires as those mental states which are not beliefs, which is, very roughly, the way the teleological argument encourages us to think about them.

5. Towards a Satisfactory Humeanism

Smith's challenge was to account for the gap that can open between an agent's sincere recognition of what she has reason to do and her having some motivation to do it. According to Smith, the problem he detected with desire-based accounts in explaining this phenomenon was a problem with such accounts *per se*. The trouble was supposed to arise once desire was said always to make an appearance among the conditions of having a

reason. For it would seem, then, that the agent must always be motivated to do what she has reason to do, since desires are intrinsically motivating states. But I do not think that there really is a problem here. That desire-based accounts do not seem to be able to account for the distinction between having a goal and being motivated to pursue it is not a function of their being desire-based. Rather, I will argue, it is a function of the notion of desire that Smith assumes they must employ. Smith thinks that desire-based accounts will take over the (very popular) conception of desire embodied in the teleological argument. Now the idea that desires are essentially mental states with a certain direction of fit might serve adequately to distinguish desires from beliefs, but it does not serve to give an adequate account of desire. It gives us only a highly schematic, conceptual account of desire. What a successful desire-based account of valuing needs in order to meet Smith's challenge is a more subtle, more psychologically informed account of desire.

The primary task, then, is to show that Smith is mistaken in his claim that desire-based accounts cannot meet his challenge. To do that, we need to find a more sophisticated account of desire than that found in the "direction-of-fit" account that Smith favors. I will very briefly sketch some aspects of the account found in psychoanalytic writings. Even if the psychoanalytic account is thought ultimately to be without merit, the case against Smith will be advanced. For our primary task is to show that Smith is wrong to hold that desire-based account are, in principle, unable to account for his challenge – that what trouble Smith sees for such accounts is due, not to their being desire-based, but to the conception of desire he assumes they will employ. Pointing to an alternative and plausible account of desire (even if that account ultimately fails) would support our claim.

To see how the direction of fit conception of desire encourages us to overlook a crucial feature of desire, consider the relation between desires and wishes. On the direction-of-fit account, both are simply mental states with the same direction-of-fit (see the discussion above, p. 165). But clearly the desire and the wish are importantly different attitudes, even if they are importantly similar. And some of the most important differences between them come out precisely in how they enter into our deliberation, how they affect or don't affect what we have reason to do, and under what conditions they move us. Wishes, for instance, do not normally make the same demands on our deliberations as do desires, even though both are most naturally construed as desire-like states as this difference is drawn via the notion of direction of fit. We need an account of desire, or desire-like states, that can make sense out of such phenomena as

described above. The account of desire embodied in the teleological argument cannot do that. For the only resource it gives us to distinguish between mental states is the notion of direction of fit. Since both desires and wishes have the same direction of fit, it cannot reveal the important differences between the states. We have spoken of the "gap" that can open between an agent's correct acceptance of the claim that she has a reason to act and her motivation to act on that reason. To focus our attention, let us present what seems to be a typical situation of this sort in more detail.

Suppose an agent sinks into a state of very low self-regard. Over a period of time she has meet with a series of minor setbacks. She has found none of these of great significance, and she certainly is not brooding over some particular event. Still, the cumulative effect of these events is a general lowering of her opinion of herself. And one effect of this is that she comes to treat her desire for advancement in her field, and all that this entails, quite differently than before. There would be some truth in saying that she starts to treat the desire as if bringing about its satisfaction were beyond her control: she treats her desire as if it were a wish. We should be careful, though. I say that she treats the desire "as if" these things were the case because it would be wrong to explain her lack of motivation by attributing to her the belief that these things really are beyond her control. These things are not beyond her ability and she knows it. Not only does she believe it is in her power to advance her career, at some level she even expects to do so; she does not see herself languishing in this state forever. However difficult it is to give an adequate account of her state, we look in the wrong direction if we attempt to explain her practical irrationality by attributing to her an additional mental state, whether in the form of a belief or a desire.

What I mean by saying she treats the desire as if it were a wish is that in some respects (though certainly not all) the role the desire occupies in her mental life is a role more appropriate to a wish. In deliberation, for example, the desire, like a wish, does not demand that action be taken to satisfy it. Rather, it enters her deliberations as something representing a possibility that might come to pass, a possibility to which she looks forward. Unlike a wish, though, it does make some demand in her deliberations: because she has this desire she plans around a successful career. When she thinks about how to organize her life over the next few years she will make assumptions that make sense only given that she works hard. So the desire is not idle; it affects her deliberation. But the desire has undergone a change in the kind of demand it makes. It no longer demands that she act toward its satisfaction. Furthermore, there

are changes in the role the desire plays outside of her deliberation as well. Before, under the prompting of the desire she would work. Now, the same desire serves mainly to bring on episodes of self-rebuke or self-pity, which in fact prevent her from acting toward what she wants.

It is important to see that she does not treat the desire as though it were not entirely hers, a foreign invader, as it were. That is an apt characterization in the familiar example of the addict who wants to rid himself of his cravings. But it is not usually an apt description of the way we treat our wishes or hopes or some of our fantasies. Our agent is still identified with her desire. She still thinks of herself as someone who is striving in the direction in which it points, and it may be important to her that others think of her in that way. Still, for the reasons mentioned, she is not actually going in that direction.

I have said that we might understand how an agent can fail to be motivated to act on the reason given her by her desires by seeing the desire as occupying a role more appropriate to a wish. The objection is that an appeal to a wish could not explain what needs to be explained. For the reason a wish is not motivational is that its object is beyond the agent's reach, or at least, the agent must believe it to be beyond her reach. And it is the presence of that belief that renders the wish idle. Indeed, it is just because of the presence of such a belief that we call a desire a wish. But that kind of belief is not present in our example. Our agent does believe that she can do what is needed to advance her career. So the very condition that seems to explain why wishes are non-motivational is absent in our case. Therefore we cannot have an explanation of our agent's lack of motivation simply by saying that she treats her desire as if it were a wish.

This objection rests on the following, common understanding of a wish. A wish is nothing but a desire that the agent thinks she can do nothing to satisfy. On this account what makes a wish different from a desire is purely an extrinsic matter: it is simply the desire of an agent who also happens to have a belief about the impossibility of doing anything to satisfy it. Therefore, this understanding of what a wish is is perfectly compatible with the direction-of-fit account of desire and of mental states generally, since no intrinsic property of the desire or wish needs to be appealed to in order to distinguish one from the other. Both can be adequately characterized as mental states with the world-to-mind direction of fit. What distinguishes then is the presence (or absence) of relevant beliefs (and these, in turn, can be characterized as mental states with the mind-to-world direction of fit). Now, there is no point in denying that we sometimes use the word 'wish' in this way, simply to

designate a desire whose object lies beyond the reach of the agent, or the satisfaction of which an agent supposes he can do nothing to bring about. But the objection presupposes that is the *only* understanding of a wish available to us, and so one way we can answer the objection is by showing that its presupposition is mistaken. We need a different understanding of the wish.

The understanding I want to consider comes to us from Freud. And in a way absolutely typical of Freud his account stands common understanding on its head. What needs to be explained is the non-motivational character of the wish, and the standard account just given traces this character to the agent's belief that there is nothing to be done to satisfy the wish. The wish cannot combine with an instrumental belief and move the agent to act, because there is no such instrumental belief. In contrast to this, Freud's account traces the non-motivational character of the wish to the fact that the agent has already done *enough* to satisfy it. It is because the wish is already satisfied that the agent takes no steps – no further steps, we might say – to satisfy it. The wish is satisfied because, on one understanding of Freud's account, when we wish we also tend to imagine the wish satisfied, and just because we imagine the wish satisfied, it is for us as though it were. That is, it is absolutely characteristic of the Freudian wish that it tends to be satisfied, at least temporarily, by the very act of imagining that it brings in its wake.

Why should it be for us that the wish is satisfied just when we imagine it to be so? A full discussion would lead us too far afield, nor, again, do the details of Freud's account matter for the plausibility of the general sort of answer I am trying to provide. But the central idea is that the wish itself, and the sort of imagining that it brings in train, both exist within the orbit of an archaic theory of mind which Freud called the "omnipotence of thought." The omnipotence of thought involves a drastic over-valuation of the power of our mental processes, and in particular an over-valuation of the power of our imagination. Under the influence of this archaic sort of thinking, we attribute to our imagination the power we normally attribute to reality: the power to satisfy our desires. Still, we need not go so far as to say that under the influence of the omnipotence of thought an agent fully believes the desire or wish to be satisfied (Hopkins 1982). That is, he need not believe that the world really does fit the content of the desire. In some cases that may be the outcome, and when it is we are in the territory of self-deception and motivated irrationality. But before we find ourselves all the way into that territory, we can attach a sense to the idea that it is for the agent "as though" the wish were satisfied if we think of the satisfaction that imagination brings

to the wish as affective, instead of cognitive satisfaction: the anxiety associated with frustrated desire is, for the moment, relieved.

The crucial feature of the Freudian account for our purposes is that desires can "regress" to the state of the wish. So, for example, a certain desire under the pressure exerted upon it by frustration, frustration which the agent finds intolerable, might find (at least partial, at least temporary) satisfaction through imaginative activity: the desire, we might say, returns to the state of a wish. Adopting this view requires us to accept two specifically psychoanalytic features of mental states (Gardner 1993, p. 123). The first is that mental states can be ranked according to their maturity. More mature manifestations of mental states evolve from more infantile beginnings, and one mark of the maturity of a mental state is its mode of satisfaction, e.g., whether it must be satisfied through action or can be satisfied through activity of the imagination.

The second feature of mental states that the psychoanalytic account requires, and this is the important point for the views I am defending, is that they can be individuated in some way other than by appeal to their conditions of satisfaction. For it is, throughout, the same mental state (once a desire, now a wish) that can find satisfaction now only in action, now also in imaginative activity. This is important for our account because it illustrates the idea that I have tried to express by talking about the roles of desire, and the differing roles that desires can play within the psychology of agents. For when a desire regresses to the state of a wish, its role has changed, and it is in terms of the change of its role that we can account for how a desire might cease to be motivational.

Again, this defense of Humeanism does not rest upon the ultimate validity of the Freudian account. The case is advanced once we see how a more sophisticated account of desire than that embodied in the teleological argument could have the resources to meet Smith's challenge. I believe the Freudian account has such resources, though there may be other accounts that do as well. The point is that once we have such an account of desire, then a central reason for rejecting Humeanism goes by the boards. Of course, one can simply deny desire *any* role in the justification or explanation of action. Nothing I have said here tells against this thoroughgoing rejection of Humeanism. But if you find it plausible that desire does have a central role in the motivation and explanation of action, then Smith's challenge gives no reason for thinking desire cannot play an essential role its justification of action as well.

Reed College
Department of Philosophy
3203 SE Woodstock Blvd
Portland OR 97212, USA
e-mail: arkonovs@reed.edu

REFERENCES

Arkonovich, S. (2001). Defending Desire: Scanlon's Anti-Humeanism. *Philosophy and Phenomenological Research* **63** (3), 499-521.

Gardner, S. (1993). *Irrationality and the Philosophy of Psychoanalysis*. Cambridge: Cambridge University Press.

Hopkins, J. (1982). Introduction. In: R. Wollheim and J. Hopkins (eds.), *Philosophical Essays on Freud*, pp. xxv-xxi. Cambridge: Cambridge University Press.

Humberstone, I.L. (1992). Direction of Fit. *Mind* **101** (401), 59-83.

Korsgaard, C. (1996). Skepticism About Practical Reason. In: *Creating the Kingdom of Ends*, pp. 311-334. Cambridge: Cambridge University Press.

Nagel, T. (1970). *The Possibility of Altruism*. Oxford: Clarendon Press.

Quinn, W. (1994). Putting Rationality in Its Place. In: *Morality and Action*, pp. 228-255. Cambridge: Cambridge University Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schueler, G.F. (1991). Pro-Attitudes and Direction of Fit. *Mind* **100** (2), 277-281.

Smith, M. (1987). The Humean Theory of Motivation. *Mind* **96** (381), 36-61.

Smith, M. (1992). Valuing: Desiring or Believing? In: D. Charles and K. Lennon (eds.), *Reduction, Explanation, and Realism*, pp. 323-360. Oxford: Oxford University Press.

Smith, M. (1995). *The Moral Problem*. Cambridge, MA.: Blackwell.

Stocker, M. (1979). Desiring the Bad: An Essay in Moral Psychology. *Journal of Philosophy* **76** (774), 738-753.

Williams, B. (1995). Replies: Internal and External Reasons. In: J.E.J. Altham and R. Harrison (eds.), *World, Mind and Ethics*, pp. 186-194. New York: Cambridge University Press.

Wollheim, R. (1984). *The Thread of Life*. Cambridge, MA: Harvard University Press.

Wollheim, R. (1999). *On the Emotions*. New Haven, CT: Yale University Press.

Carla Bagnoli

PHENOMENOLOGY OF THE AFTERMATH: ETHICAL THEORY AND THE INTELLIGIBILITY OF MORAL EXPERIENCE

"Theory is needed to refresh the tired
imagination of our practice"

Iris Murdoch
A House of Theory

It is a matter of contention whether we need ethical theory at all. Critics argue that ethical theory does not serve any genuine purpose, and urge us to resist the temptation of theorizing when reflecting about morality (see Williams 1985, pp. 93-119; Baier 1985; Noble 1989, pp. 49-64). According to Bernard Williams, ethical theory is not simply irrelevant, but a misguided enterprise, whose primary purpose is not to understand our experience, but to criticize it, correct it, and explain it away (see Williams 1965, Baier 1985). On what grounds, he asks, does ethical theory have the authority to do so?

This question has become the main focus of current debates about the aims and ambitions of ethical theory.[1] Some argue that the appeal to moral experience represents independent and autonomous evidence that suffices to counter ethical theory. Others rebut that moral experience should be interpreted on the basis of theoretical considerations, and thus cannot be used as autonomous evidence to assess the viability of ethical theory (MacIntyre 1990; Donagan 1996; McConnell 1996, pp. 36-47; Morris 1992).

[1] On the debate about how to understand the agent's experience in the aftermath of moral conflict, see also Barcan Marcus (1980), Gowans (1996), Mothersill (1996, pp. 66-85). Reconstructing the debate over the possibility of moral dilemmas, Gowans talks of the difference in style of moral reflection and labels the two groups respectively as "experimentalists" and "rationalists," see, e.g., Gowans (1996).

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 185-212. Amsterdam/New York, NY: Rodopi, 2007.

In this paper, I will show that it is misleading to focus on the issue whether ethical theory or phenomenology is supremely authoritative. First, this approach misrepresents from the start the relation between theory and moral experience, suggesting that there is a gap between the theory and the practice of morality. Second, it misconstrues the anti-theory critique and encourages replies that are partial or question begging. The point of this critique is to suggest that there is an alternative between theory and mere prejudice, namely, a reflective stance. The issue is whether theorizing is conducive to our understanding of moral experience; and defenders of ethical theory still bear the onus of proof.

My argument is meant to refocus the debate over the viability of ethical theory by revisiting the claims about the nature of theorizing. I will argue that theorizing in ethics is in itself a moral activity, continuous with our moral practices, and meant to further our understanding of the experience and aspirations we have. The distinctive purpose of theorizing is to propose a plausible and decent ideal of moral agency. In assessing the viability of ethical theory, we should consider whether it offers an intelligible picture of ourselves and posits challenges that it is worthwhile for us to undertake. On the basis of this conception of theorizing, I argue that moral phenomenology represents a test of adequacy for ethical theory to the extent that it imposes on it a requirement of intelligibility. Appeal to the agent's experience is therefore used not as a basis to counter ethical theory, but to set its agenda.

1. Williams' Argument from Moral Phenomenology

Williams' critique of the aims and ambitions of ethical theory starts with some considerations about a particular moral phenomenon, namely, the experience of moral conflict. In a series of seminal papers, he argues that the agent's experience of regret reveals the possibility of intractable moral conflicts, and shows that the very enterprise of ethical theory is hopelessly misguided (Williams 1963; Williams 1965; Williams 1973a; Williams 1981, pp. 54-70).

A moral conflict is a case in which "there is a conflict between two moral judgments that a man is disposed to make relevant in deciding what to do" (Williams 1963, p. 170). Agamemnon is representative of this kind of predicament, which is further qualified as a conflict of obligations. Agamemnon's thinks that he ought to sacrifice Iphigeneia,

but this conviction does not relieve the anguish for failing his most basic parental duty. To make sense of Agamemnon's regret, Williams argues, we have to recognize that his deliberation left a remainder. More precisely, Agamemnon's regret proves that he is still bound by his obligation to Iphigeneia, despite the fact that he deliberated and acted against such obligation (Williams 1963, pp. 173-174).[2]

According to Williams, this remainder is an important feature of moral conflicts. The agent who is bound by conflicting moral claims does not ask herself: "I seem to have two incompatible duties, hence I must be mistaken about one of them. What am I missing?" In thinking about what to do, this agent is not attempting to expunge a seeming duty, but to respond to both moral claims.[3] For Williams, this marks the difference between the structure of moral conflicts and the structure of conflict of beliefs. In the case of conflicts of beliefs, the discovery of the conflict initiates a coherence-driven revision aiming at banishing the error. But moral conflicts are importantly different also from conflicts of desires in which the agent might simply decide to ignore one item and act upon the other. The possibility of freeing oneself from moral conflict by withdrawal is precluded by the very idea of a moral claim,[4] which demands to be recognized and acted upon or else it resurfaces in another guise, (e.g., in the guise of regret). This implies that any attempt to resolve moral conflicts by rejecting the authority of one of the claims at stake is doomed to failure: "moral conflicts are neither systematically avoidable, nor all soluble without remainder" (Williams 1963, p. 179). Consequently, to think of ethical theory as providing a technique to eradicate moral conflict is to misunderstand the nature of moral claims and falsify the logic of moral thought (Williams 1963, p. 183).

This conclusion is certainly at odds with the deliverances of many ethical theories, which aim at resolving moral conflicts by annulling the authority of the overridden ought. Williams' point is that phenomenological considerations suffice to counter such theories.[5] In

---

[2] Williams writes: "The item that was not acted upon may persist as regret" (1963, p. 183). As Williams remarks elsewhere, there might be other expressions of the remainder: the agent might undertake action, offer an apology, a deal, or an explanation, see Williams (1981, p. 74).

[3] Williams writes: "I do not think in terms of banishing an error. I think, if constructively at all, in terms of acting for the best, and this is a frame of mind that *acknowledges* the presence of both the two *ought's*" (1963, p. 172).

[4] "The notion of moral claim is of something that I may not ignore hence it is not up to me to give myself a life free of conflict by withdrawing my interest from such claims" (Williams 1963, p. 178); see also Williams (1981, p. 75).

[5] "It is a fundamental criticism of many ethical theories that their account of moral

"Ethical Consistency" and other early essays, this criticism is explicitly directed against moral realism, which treats moral judgments as assertions and therefore adopts an epistemic model of rationality. For this reason, Williams insists on the asymmetry between moral conflicts and conflicts of beliefs. However, to understand the full force of Williams' argument, we should situate it in the broader context of the dispute about the nature and scope of deliberation and the authority of ethical theory. In *Ethics and the Limits of Philosophy*, it becomes apparent that for Williams ethical theory as such is a machinery devised to get rid of moral conflicts just in the way that runs against our experience of it. Ethical theory, as Williams defines it, is mainly concerned with designing a general test for the correctness of moral beliefs. He writes:

> Here the aim of a theory is not simply, or even primarily, to understand conflict. We have other ways, historical and sociological, of understanding it. The aim of a theory is rather to resolve it, in the more radical sense that it should give some compelling reason to accept one intuition rather than another. The question we have to consider is: How can any ethical theory have the authority to do that? (Williams 1985, p. 99)

Because of their concern for tests, ethical theories tend to answer this question starting from just one aspect of ethical experience, that is, moral beliefs.[6] Other factors have concurred to shape ethical theory this way. For example, Williams alleges that the predominance of Kantian rationalism and analytic philosophy have been especially detrimental to our understanding of these issues in that they have caused a pernicious neglect of moral psychology. Apart from the credibility of these allegations, it is apparent that Williams' critique takes aim at all varieties of ethical theories (whether realist or not).

Defenders of ethical theory have pursued two lines of defense. First, they have argued that whether emotional experience is morally significant is an open question to be resolved against the background of ethical theory.[7] Bare reference to what the agent feels proves nothing

---

conflict and its resolution do not do justice to the facts of regret and related considerations: basically because they eliminate from the scene the ought not acted upon" (Williams 1963, p. 175).

[6] "The ethical theorist tends to assimilate conflicts in moral beliefs to theoretical contradiction, and applies to moral understanding a model of theoretical rationality and adequacy" (Williams 1981, pp. 80-81). On the ambiguity of the notion of moral experience and the theorist's focus on beliefs, see also Williams (1985, p. 93; 1973a, pp. 166, 207-229).

[7] I discuss this reply in the next section.

about ethical theory. Second, they have argued that ethical theory belongs to our common practices, and to this extent, it serves some genuine purposes.[8] In my view, these are only partial replies to Williams' critique because they take for granted that we need ethical theory for reflecting on our practices: this answer cannot satisfy the anti-theorist. My argument is meant to refocus the debate over the viability of ethical theory by reexamining the nature and purpose of theorizing in ethics. By this route, I will show that Williams' argument does not have the destructive impact on ethical theory commonly alleged.

2. The Appropriateness of the Agent's Experience

Williams' most general contention is that there is something in the nature of ethical theory that leads us astray when reflecting on our own experience. More precisely, ethical theory aims at correcting, reforming, and to this extent, discounting the agent's experience. Defenders of ethical theory may just say that this is both inevitable and welcome[9] in that ethical theory is necessary to determine the relevance, significance, and appropriateness of our experience. This task seems to be especially worthwhile when we think of our emotional experience. Consider, for example, the case in which one regrets having to decline a dinner invitation. How does this case differ from Agamemnon's regret? And how do we establish that in one case the feeling does not have moral status, while in the other case it does?

Starting from these considerations, one can object that Williams' argument from regret is inconclusive because the criteria of appropriateness of regret are so weak (or so diverse) that they apply also in cases of non-moral failure (e.g., breaches of etiquette). A plausible rejoinder could hardly invoke varieties of feelings that are markedly moral, like remorse or guilty feelings.[10] It is easy enough to imagine cases where one experiences remorse or guilty feelings even if there is no moral failure. For example, as a result of a strict religious education, Ingrid feels guilty for having sex out of wedlock, although she does not take herself to be violating any moral obligation. Holocaust survivors are often reported to feel guilty for having survived, and many rape victims

---

[8] This is Peter Railton's view, for which I account in section 4. See also Gibbard (1995).

[9] As I take it, this is Gibbard's line of reply, Gibbard (1995).

[10] See, e.g., Barcan Marcus (1980). Some philosophers focus on guilty feelings as samples of a larger category, see, e.g., Greenspan (1994). For an alternative view, see Foot (1995), Fingarette (1979).

confess to blaming themselves because they think they deserved or caused the violence.[11]

It seems that to understand these phenomena we need some normative criteria for evaluating their moral relevance, significance, and appropriateness. To raise questions about the moral significance and appropriateness of emotions is to recognize that the agent's experience of the situation at hand is and should be critically assessable. The issue to settle here is not whether ethical theory should be more authoritative than the agent's experience, but to rethink what it is to take the agent's experience at face value.

In this vein, Richard M. Hare argues that ethical theory should account for moral phenomenology, but does not have to take the agent's experience as veridical or supremely authoritative. In the face of moral conflicts, for example, ethical theory appears to have two major objectives. Its first normative task consists in offering to the agent some criteria to evaluate her situation, determine her obligations and consequently assess whether her feelings are appropriate. Its second task is explanatory and consists in uncovering the psychological mechanisms that account for the reasons why the agent perceives the situation in this given way (whether in accord or not with ethical theory). In case of moral conflict, for example, the agent is taught how to resolve her issue and explained why she is under the false impression that there is no resolution to be found.

Suppose that Beatrix correctly deliberated to the conclusion that she ought to assist her ailing mother rather than attend a friend's exhibition as promised, but experiences guilty feelings for having broken her promise. This is an example of a successful pattern of deliberation, whose conclusion is an all-things-considered judgment stating an overriding reason. Ethical theory fulfills its normative task when it provides the agent with the canons of moral reasoning. But in Beatrix's case, ethical theory has the further task of explaining the reasons why Beatrix experiences guilty feelings even though she is confident that she did the right thing. The upshot of this investigation is not that the agent's feelings are misplaced and carry no moral relevance, but that they do not mark any failure in deliberation. This finding is morally significant; for example, it tells us that Beatrix's feelings do not represent a ground for

[11] Cases of collective guilt are very controversial because of the absence of imputability and intentionality, see, e.g., Foot (1995, p. 117), Gibbard (1990, p. 137), Greenspan (1995). Van Fraassen holds that guilty feelings are misplaced if the agent did not have a better alternative. Morris treats them as non-moral cases, see Morris (1992, pp. 223-227).

reexamining her decision, but perhaps for undertaking some further reparatory actions.

Hare's explanation of Beatrix's experience of guilty feelings is based on the idea that ethical theory should distinguish between perfect and imperfect levels of moral rationality. Perfectly rational beings are capable of reasoning and acting on the basis of act-utilitarianism. Imperfect rational beings like us, lacking the cognitive and logical capacities to operate according to act-utilitarianism, should reason on the basis of general rules justified on utilitarian grounds. Moral rules provide us with general and thus defeasible directions as to what to do, that is, *prima facie* duties that can be overridden by deliberation (Hare 1981, p. 39). When we learn the content of these general rules, we also learn to feel bad at not accomplishing our duty.[12] Guilty feelings are thus attached to *prima facie* duties, and for this reason, when a *prima facie* duty is overridden one is likely to feel bad, even if one deliberated correctly. This explains Beatrix's experience: her guilty feelings survive deliberation even though the duty to which they are associated does not.[13] On this account, then, Beatrix's guilty feelings are not to be considered genuinely residual because they do not signal a loss of value or a failure in deliberation; nonetheless, they represent a rational and appropriate response. This is because a *prima facie* duty is not a seeming reason that is eventually trumped by the real moral duty. An all-things-considered judgment does not simply restate one *prima facie* duty of the two in conflict: its normative status is different than the prima facie duty because it is the conclusion of a deliberative process through which

[12] On Hare's view, internalization is therefore a crucial step in learning to reason morally and also accounts for the mechanism of moral motivation. If we were not able to represent moral rules internally, moral motivation would become a completely mysterious phenomenon. For a similar account of the role of feelings in the explanation of moral motivation, see Gibbard (1990, pp. 68-71, 75-76). According to Gibbard, habituation by internalization of moral rules warrants the agent's integrity and social reliability. Guilty feelings measure individuals' compliance and shape mutual expectations. Contrary to Gibbard, Hare does not pay much attention to the social function of internalization of rules.

[13] On this picture, the decent moral agent is taught and expected to feel guilty even when she is not at fault, and one might object that this portraits morality as a too punitive practice, hardly justifiable on prudential grounds; see, e.g., Statman (1995, pp. 47-51). Like any utilitarian philosopher, Hare can avail himself to a prompt reply to this objection. Although apparently dysfunctional, a system of morality where moral agents are taught and expected to feel guilty when they violate *prima facie* duties (and *a fortiori* when they fail to fulfill a duty) is overall justified, that is, justified on the basis of utilitarian considerations.

competing *prima facie* duties are assessed.[14] Moreover, all-things-considered judgments do not issue a new ranking or a revision of the ordinary moral rules that the agent adopts, and therefore do not cancel the normative force of overridden *prima facie* duties.[15]

This explanation of the emergence of guilty feelings in case of successful deliberation does not require the agent to reclassify her experience. Thus, Hare's proposal qualifies as an interesting alternative to Williams' because it does not discount the agent's experience and does not make her conflicts adventitious.[16] The all-things-considered judgment settles the conflict at a given time, but it does not prevent the possibility that a conflict of the very same kind may arise in the future. This is the result of having acknowledged that all-things-considered judgments differ in normative status from *prima facie* duties.

Regretfully, little attention has been paid to the nature of this distinction, which is in fact crucial to locate exactly the disagreement between Williams and Hare. They concur that all-things-considered judgments and each of the conflicting *prima facie* duties differ in normative status, and that the all-things-considered-judgment does not merely replicate one of the *prima facie* duties. For Hare, an all-things-considered-judgment represents the moral resolution of the conflict; and while such resolution might bear a cost, which can be marked by guilty feelings, it cannot leave a moral remainder. For example, if Agamemnon reached the all-things-considered judgment that he ought to sacrifice Iphigenia, his regret is morally appropriate but it does not show that deliberation failed. Rather, it shows that he recognizes that his parental duty is still morally important. To the contrary, Williams argues that in case of moral conflict, deliberation leaves a moral remainder even when it yields an all-things-considered-judgment about what to do because all-things considered judgments are not moral resolutions.[17] Ultimately,

[14] In insisting on this distinction, Hare proves to be a charitable reader of D.W. Ross. Ross emphasizes that the qualification '*prima facie*' should not be taken as suggesting that it is only an apparent reason, but "an objective fact involved in the nature of the situation" (Ross 1930, p. 20).

[15] This is because if we were to build all the exceptions in the body of the rules, such rules would become so complex that it would be impossible for us to learn and use them.

[16] Williams acknowledges this much, see Williams (1963, p. 176). His qualms seem to be confined to Hare's conception of what constitutes a resolution for a moral conflict. In this case, then, the objection concerns Hare's utilitarianism rather than his conception of *prima facie* duty.

[17] Williams (1963, pp. 184-185). Williams holds that is a mistake to conflate the status of all-things-considered judgments (which is a deliberative judgment answering the deliberative question "What ought I to do?") from moral claims (like *prima facie* duties are) for reasons independent of the present issue. He believes that this conflation leads

Williams and Hare disagree about the moral status of all-things-considered judgments, and whether they count as moral resolutions (as opposed to a merely deliberative way out).

The question to address here is whether in order to vindicate moral phenomenology, we have to admit that Agamemnon's regret signals a moral remainder. This does not seem obvious. To be descriptively plausible, an ethical theory must explain the case in which the agent can feel guilty feelings even though she deliberated successfully, and reached an all-things-considered judgment. As Hare argues, the all-things-considered judgment counts as a moral resolution, but it does not cancel the moral significance of each of the *prima facie* duties in conflict. Therefore, negative feelings are rationally justified even when there is a moral resolution to one's conflict, as it happens in Beatrix's case. Hare's point is that this much is enough to accommodate moral phenomenology in that it offers a plausible account of the agent's experience, which does not require any radical revision or reclassification.[18] That is, the distinction in normative status between all-things-considered judgments and *prima facie* duties is sufficient to fully vindicate moral experience. It is not necessary to cast such distinction as a distinction in moral status, as Williams does.

If Williams were to claim that the phenomenology of moral conflicts proves ethical theory to be descriptively inadequate, then he seems to have missed the target. Hare's theory is an example of descriptive adequacy. However, as I have suggested, there is another and more interesting claim that Williams is pressing. To represent morality via ethical theory is not just unrealistic but also detrimental to our

---

one to defend the supremacy and ubiquity of morality, since the deliberative question can be asked whether or not the claims in question are moral. But it is the presumption that deliberative judgment cannot be qualified as moral that allows him to say that Agamemnon did not resolve his moral problem because he could not reach a moral resolution that canceled the authority of one of the moral claims at stake.

[18] The language of *prima facie* and all things considered duties is Ross', and Williams remarks: "Ross makes a valiant attempt to get nearer to the facts than this, with his doctrine that *prima facie* obligations are not just seeming obligations, but more in the nature of a claim, which can generate residual obligations if not fulfilled. But it remains obscure how all this is supposed to be so within the general structure of his theory" (Williams 1963, p. 176). See also Williams (1981, p. 73), where Williams rejects the terminology of *prima facie* duties as "ambiguous." On neither occasions does Williams make clear the reason why such terminology cannot be used exactly to make his own case. Hare does not face the difficulties that Williams points out in Ross because his account of internalization explains moral motivation in a naturalistic way; in this respect, it is very similar to Gibbard's.

understanding of ourselves and the practice of morality. This is the claim we must address next.

3. Accounting for Moral Phenomenology

Recent debates about the importance and role of negative feelings in the aftermath of moral conflict have taken this shape: *Either* we take the agent's moral experience (e.g., of guilty feelings) as a standard for assessing the viability of ethical theory, *or* we take ethical theory to provide the criteria for judging whether such experience is appropriate. Some take Williams to argue that we can appeal to moral phenomenology as a fact against which we ought to test the plausibility of ethical theory.[19] To this claim, others reply that we should account for moral phenomenology from within ethical theory because the mere appeal to the agent's experience is hopelessly circular (MacIntyre 1990, p. 367; Donagan 1996, p. 11; McConnell 1996, pp. 37-39). We need guidance in interpreting the agent's moral experience, and we need to assess whether the agent's feelings are appropriate; such criteria for interpretation and assessment are dictated by ethical theory. To overcome this stalemate in the debate about methodology, I propose that we reconsider the relation between ethical theory and phenomenology afresh.

We can make some progress by elucidating the very notion of moral phenomenology. Many follow Williams in referring to "the facts of regret and related considerations" as part of moral phenomenology (Williams 1963, p. 175).[20] As MacIntyre remarks, it is not clear to what conception of "facts" these philosophers avail themselves (1990, p. 377).[21] If we take

[19] Barcan Marcus agrees with Williams that to deny that there are grounds for guilty feelings in cases of moral conflict is "false to facts," but she shows that this does not bear the consequences that Williams alleges concerning consistency, see Barcan Marcus (1980). Gowans agrees with Williams, although he argues for a qualification of the appeal to moral experience, see Gowans (1996).

[20] See also Barcan Marcus (1980), and Gowans (1996). In referring to the guilty feelings that characterize the agent's experience of moral conflicts, Van Fraassen talks of "the kind of facts of moral life on which ethical theories founder" (1973, p. 17). Railton writes: "Pluralism and dilemma come on the scene as purported facts of moral experience – and who can wonder?" (1992, p. 720); later on, he speaks of "hard-to-ignore intuitive anomalies" and "recalcitrant phenomena" (1992, p. 723).

[21] On the same point, see also Sinnott-Armstrong (1988, p. 34). This objection stands even when one appreciates the value of literature and narrative reconstruction. In order to make a case for the relevance of literary and autobiographical remarks, we need to offer arguments about the nature of practical deliberation, and the role of examples. See, e.g., Dancy (1985).

phenomenology to be constituted by "brute facts" about how the agent feels, facts that are said to be independent of any theoretical consideration, there are an overwhelming amount of reasons for rejecting Williams' conclusions. When based on "brute facts," arguments from experience do not prove (or disprove) ethical theory. These arguments are inevitably incomplete: they work only in combination with other presuppositions (some of which are theory-dependent, e.g. those concerning the canons of moral reasoning), and ultimately do not tell us anything interesting about the powers and capacity of ethical theory.

While I agree with McIntyre that these arguments are inconclusive, I also think that to frame the dispute in terms of brute facts is to misunderstand the very idea of moral phenomenology and, consequently, its bearing on ethical theory. Both parties in this dispute are at times guilty of an equivocation about the notion of moral phenomenology, and to my view, nobody has provided a satisfactory account of what moral experience is supposed to be and to show. Once moral phenomenology is understood correctly, the issue of what one's experience really shows about ethical theory becomes specious.

We would do a disservice to moral phenomenology if we suggested that the agent's moral experience is composed of brute facts. Moral experience results from a thick and intricate web of complicated relations between one's sensibility, one's moral outlook and practices, and one's assessment of the situation. What the agent takes her moral experience to be depends on the concepts and the language of morality which she finds available and which she understands. For example, Beatrix's guilty feelings are experienced as the perception of a moral failure insofar as she comprehends what it is to break a promise. Whether such feelings are appropriate or not is a question that arises exactly because the significance of these feelings depend on theoretical considerations concerning, for example, the canons of moral reasoning, background conceptions of moral competence and agency, criteria of salience, and normative criteria for assessing the situation. In order to understand the significance of the agent's experience, and their impact on ethical theory, we are bound to rely on other considerations beside the agent's reported emotions. First, we need criteria for interpreting the agent's experience so that it makes sense from within the agent's own narrative. This amounts to reconstructing and making the agent's experience intelligible from the agent's own perspective. Second, we need criteria for justifying this experience morally. Both kinds of criteria are normative, but the task of making something intelligible is not the same as the task of justification. There might be several kinds of alternative reconstructions

that make that provide a coherent narrative structure for the agent's experience. And there might be several ways of justifying. The point here is simply that the question arise whether the agent's experience is justified arises only after the issue of interpretation and intelligibility has been settled. Since interpretation involves many kinds of theoretical considerations, it is misleading to conceive of this debate as a dispute about whether the agent's experience is prior to and independent of theoretical considerations. It clearly isn't.

However, Williams hardly suggests that in accounting for moral experience we are dealing with brute facts. While he urges that philosophy should be responsive to experience, he remarks that: "At the same time it is of course true that such experiences need interpretation in terms of general ideas about the status of moral thought" (Williams 1981, p. 75). His claim is that we should not be confident that there is room for separating the agent's "raw" feelings from his moral considerations.[22] Feelings entertain deep and varied relations with the agent's thoughts, habits, and beliefs. It is because feelings can be the expression of moral considerations, and are not merely attached to them, that they serve to assess the viability of ethical theory. The relations among the moral feelings and the explicit or implicit moral considerations may be diverse, and one should not suppose that feelings have only the purpose of expressing moral beliefs.[23] But it is important to notice that to address the question of moral phenomenology is not to ask whether the agent's feelings square with her beliefs about the situation. We are dealing with something much more complex. To appreciate this complexity is also to realize a variety of normative aims that ethical theory is supposed to

[22] "Are we really to think that if a man (a) thinks that he ought not to cause needless suffering and (b) is distressed by the fact of the prospect of causing needless suffering, then (a) and (b) are just two separate facts about him? Surely (b) can be one expression of (a), and (a) one root of (b)? And there are other possible connexions between (a) and (b). If such connexions are admitted, then it may well appear absurdly irrealistic to try to praise apart a man's feeling regrets about what he has done and his thinking that what he has done is something that he ought not to have done, or constituted a failure to do what he ought to have done" (Williams 1963, p. 174). As it appears, Williams is anticipating Hare's move here.

[23] For the purpose of this argument, I insist on the expressive role of feelings, but I believe that they also play evaluative and deliberative roles in our moral life. On some occasions, they are moral responses, such as when they mark the perception of a moral failure and the need for apology or reparation. This is the expressive function I mentioned in the text. Generally, emotions are ways in which we are sensitive to the salience of some traits of the world. In this sense, they shape and guide our practical reasoning. Finally, feelings are also modes of valuing, that is, ways in which we attach value to objects; this is the case of love and respect. See Bagnoli (2003, pp. 483-516).

accomplish. On this view, the agenda for ethical theory should include the investigation of the various sorts of relations between feelings, attitudes, desires, conditions of agency, and moral judgments. In this perspective, any reconstruction of the phenomenological argument as starting "from facts" misses the point about the importance and role of moral experience.

However, even admitting that moral phenomenology is infused with or organized through moral concepts, and laden with moral conceptions, it could still be at odds with ethical theory. In this event, is ethical theory more authoritative than the agent's experience? To answer this question, we are suggested to examine whether the reasons for re-describing one's experience outweigh the reasons for treating the experience as a counter-example to the theory. According to MacIntyre "how such reasons are to be outweighed will depend upon further theoretical considerations" (MacIntyre 1990, p. 371). But the question is, again, from where these "further theoretical considerations" derive their authority. There seem to be only two options. If we take these criteria to be internal, it is the agent's moral outlook that provides the ground for determining whether her emotional experience is appropriate. But this strategy seems hopelessly circular because at times the only consideration available to the agent is that very emotional experience. For example, suppose that, during a conversation, Astrid almost inadvertently shows a patronizing attitude toward her new younger colleague, and then she feels a kind of discomfort (pangs of conscience, as it were). Through the experience of this discomfort she comes to the resolution that she has to apologize to her colleague. In this case, the perception of a moral failure is tantamount to a feeling of discomfort, and cannot be tested against any other moral considerations.

If we take the criteria to be external, the agent's experience is represented in the light of the claims of a certain ethical theory, whether or not the agent endorses such a theory. Take these two examples. Ingrid feels guilty for having sex out of wedlock even though she does not regard herself as violating any moral obligation. Matilda is a rape victim tortured by feelings of guilt and self-blame. Suppose that according to a given ethical theory Matilda's sense of guilt turns out to be misplaced because it increases the amount of suffering in the world and has no reparatory function, whereas Ingrid's guilty feelings are justified because they work as deterrents against promiscuity and cement social stability. Being misplaced, Matilda's feelings are not worthy of further investigation, and should be dismissed altogether. In either case, the assessment of the significance of these feelings is established

independently of any considerations that Ingrid and Matilda allege. It could be that ethical theory finds agents to be systematically mistaken in responding with guilty feelings when they have not failed a moral obligation.[24] Such a verdict cannot be overturned simply by noticing the agents' protests to the contrary, or by observing that they knowingly persist in their mistake.

It appears that if we reconstruct the issue in terms of two competing sources of authority for the criteria of justification we are forced into a dilemma. If we endorse internal criteria of appropriateness, the challenge is to find the way of vindicating moral phenomenology without incurring the charge of circularity. If we endorse external criteria of appropriateness, the challenge is to give moral phenomenology its due. In the next sections, I propose an alternative way to address Williams' objection against ethical theory and overcome this apparent deadlock.

4. Theorizing as a Moral Activity

In the previous section, it is established that in order to interpret the agent's moral experience we ought to refer to some theoretical considerations. Critics of ethical theory, such as Williams and Annette Baier, agree that in order to reflect on one's experience some theoretical considerations are necessary, but contend that ethical theorizing is not. On their view, ethical theory is not only dispensable in order to reflect fruitfully on one's experience, but it also undermines our understanding of such experience. My aim in this section is to respond to this charge by defending theorizing as a moral activity.

The task I am undertaking is not to be understood as a defense of any particular ethical theory, but of the very activity of theorizing in ethics. In support of moral realism and Kantian rationalism, the most explicit targets of Williams' critique, further specific arguments can be and have been advanced, and I will not rehearse them here.[25] It has become apparent that Williams' (and the anti-theorists') concern is not just with particular styles of theorizing: any ethical *theory* is by its own nature at odds with moral phenomenology.

---

[24] The category of moral failure is much broader and more inclusive than the failure to fulfill a moral obligation.

[25] The literature is extensive on both subjects. In defense of realism, see, e.g., Foot (2002). In defense of Kantian rationalism, see, e.g., Korsgaard (1996), Hill (1996), Herman (1993, Chs. 7-9).

A plausible place to start in responding to Williams' objection is, therefore, his account of ethical theory:

> An ethical theory is a theoretical account of what ethical thought and practice are, which account either implies a general test for the correctness of basic ethical beliefs and principles or else implies that there cannot be such a test. (1985, p. 72)[26]

This characterization is very narrow as there are several varieties of ethical theory that clearly do not fit this characterization. Once we reform this definition, Williams' argument proves to carry a much less weight than he anticipated.[27] But my aim is not to show that some varieties of ethical theories survive Williams' attack. Rather, I want to make the case that critics of ethical theory misconstrue the shape and features of ethical theory because they misconceive of the nature and purpose of theorizing: it is this issue that calls for further investigation.

It is remarkable that in responding to Williams's critique philosophers have not directly addressed Williams' distrust of theorizing. They have generally adopted the strategy of clearing some particular normative theories from the charge of being unrealistic (Gibbard 1995). A notable exception is Peter Railton who has insisted on the continuity between moral theory and moral practice, by arguing that ethical theory is the fairly common practice of reflecting about our practices, rather than merely a philosophical aberration or contrivance (Railton 1991, pp. 185-190). However, these remarks do not yet provide a full reply to Williams' objection because they are based on the assumption that ethical theory is necessary to guide our reflection, which is exactly what Williams denies:

> [. . .] It is quite wrong to think that the only alternative to ethical theory is to refuse reflection and to remain in unreflective prejudice. Theory and prejudice are not the only possibilities for an intelligent agent, or for philosophy. (Williams 1985, p. 112)

Like Williams, other anti-theorist philosophers agree that in order to investigate our moral practices and experience we ought to deploy several kinds of theories and theoretical considerations, but argue that nothing like an ethical theory is required to this investigation.[28] For

---

[26] Annette Baier offers a similar definition, see her (1985, p. 232).

[27] This case has been very successfully made by Scanlon (1995), (1992).

[28] "There is reflection that asks for understanding of our motives, psychological or social insight into our practices, and while that may call for some kinds of theory, ethical theory is not among them. Nor is it merely that this kind of reflection is explanatory, while that which calls for ethical theory is critical" (Williams 1985, p. 112). Annette Baier raises the

example, the inquiry about the motives of a racist may require us to adopt psychological, biological, and sociological theories, but in order to identify and expose his mistake or induce him to correct his beliefs, there is no need to invoke anything like ethical theory. What goes wrong with the racist is best understood by inquiring about the force of bad habits, self-deception, false beliefs, and social deceit, rather than claiming that he resists the drive of a given normative ethical theory (Williams 1985, p. 116).[29] Similarly, if we want to understand moral conflict, we should rely on historical and sociological theories, rather than deploying ethical theory. Ethical theory contributes nothing to the understanding of moral conflict because its only aim is to resolve it.[30]

Not only is ethical theory irrelevant and dispensable to account for moral phenomena. These critics' ultimate contention is that ethical theory also inhibits our understanding of the kinds of agents we are and distorts our experience of morality. To prove this point, Williams asks us to consider the kinds of needs ethical theory is supposed to meet, and suggests that there is only one question that ethical theory is designed to answer: whether our practices are ultimately justified. The quest for justificatory reasons, being so unilateral, leads naturally to ethical theory (Williams 1985, p. 112).[31] In the attempt at responding to this demand, one inevitably commits oneself to a search for simple, general, and universal verdicts.[32] Born out of this simplification, ethical theory cannot but misunderstand and misconceive the complexity of our moral experience and practices. To be fruitful, a reflective investigation of moral life should shy away from the ambitions of ethical theory, and be wary of its alleged desiderata such as simplicity, generality, and universality (Williams 1985, p. 116).

---

same doubts about the need for ethical theory: "Moral agents also need theories, or rather the reliable facts good theories produce – facts about the way people react, about the costs and consequences of particular ways of life, on those who adopt them and on their fellow persons. We need psychological theories and social theories, and, if we are intent on political change, theories about political power and its working, and about economics. But do we need normative theories, theories that tell us what to do, in addition to theories that present us the world in which we are to try to do it?" (Williams 1985, p. 233).

[29] "This kind of irrationality is not exposed or cured by invoking, but by getting him to reflect on what he is doing. As before, this may well require some theoretical understanding of other kinds, and will involve other values" (Williams 1985, p. 116).

[30] "The aim of ethical theory is not simply, or even primarily, to understand conflict. We have other ways, historical and sociological, of understanding it" (Williams 1985, p. 99).

[31] This is also Annette Baier's conviction, see her (1986, p. 538).

[32] "Where a reason is demanded for a given practice of reason-giving, the range of possible answers gets much narrower" (Williams 1985, p. 113).

Williams seems to believe that any ethical theorist is committed to external criteria of justification, which rationalize and discount the agent's experience. Maybe the agent holds an alternative ethical theory; most likely, she does not have any theory at all. In any case, the philosopher's job is that of registering discrepancies, and noticing how much the agent's experience differs from the recommended theory. The paradigm of this kind of ethical theorist is the utilitarian philosopher intent on calculating whether and how much agent's regret affects the overall amount of utility. To such a theorist it matters what considerations the agent provides only insofar as they influence the calculation. But if the agent's experience does not have a significant impact on such a calculation, it would be hard even to justify a philosophical inquiry about it. For example, if Ingrid or Matilda's feelings decrease the overall utility, the theorist is to say that these agents, having carried their deliberation successfully, are being simply irrational and their guilty feelings are utterly misplaced. The question does not arise as to what these feelings mean to agents like Ingrid or Matilda. When the agent's judgment and experience distance from the results reached by the theory, it is always the agent who is at fault. Even when the theory systematically faults the agents' experience, this gives no pause to the theorist: He is not answerable to the agent, and recognizes no authority to moral phenomenology. His verdict cannot be overturned or undermined by observing that the agents appear to be always mistaken in their experience, or by pointing out that his theoretical reconstruction makes no sense to such agents. Williams targets exactly this kind of theorist, and I suspect that his strenuous defense of the primacy of moral experience rests on the conviction that the theorists must endorse external criteria of justification and pose as a kind of expert whose job is to expose and reform the agent's distorted vision of how things stand.

The defender of ethical theory bears the onus of proof: she should show that ethical theory addresses genuine needs, and respond to them appropriately. The thrust of the anti-theory argument is that there are no genuine needs to which ethical theory answers. Rather, ethical theory creates the fictitious need for general, simple, and universal principles because it is the result of having adopted a theoretical model of rationality. But it is not obvious why this should be so. The features of generality and universality may have emerged and selected as desiderata for practical reasons. For example, it is reasonable to speculate that there

is practical pressure to search for simplicity, generality and universality both in science and in ethics.[33]

Interestingly, Williams acknowledges this possibility when he reconstructs the rationale for the appeal of contractualist theories (Williams 1985, p. 99). Suppose there are some groups of people committed to living together peacefully, on the basis of agreement. They see this as a task of living upon principles that everybody can share, and thus search for a public procedure for the justification of their practices. Williams admits that given these assumptions, it is reasonable for these people to aspire to ethical theory. Then, he pauses to consider how demanding and distant from reality these assumptions are. What concerns me here is Williams' admission that the need for ethical theory (with features such as a public procedure) might have arisen not because one had imported a theoretical model of rationality in the practical domain, but because of eminently practical concerns and considerations (such as the task of living together peacefully on the basis of agreement). No matter how strong or whether acceptable the assumptions underlying this project are, it is apparent that their nature is normative. In fact, they would be better qualified as a moral ideal. In this case, we should explain the authority of ethical theory as depending on the appeal of a given moral ideal. One might have reservations or objections against this moral ideal, but such reservations and objections would still be motivated by practical reasons. The fundamental issue is whether this is a decent and plausible moral ideal. Such issue is normative and cannot be settled but via substantive discussion.

This brings us back to the question: Is there any reason to devise an ethical theory? That is, is there any genuine need to which ethical theory alone can answer? When confronting these questions, the ethical theorist might conceive of her job very differently from how Williams pictures it to be. She might think of ethical theory as a systematic account of practical rationality for us to use, a tool to better navigate the world. On this view, the aim of ethical theory is to assist the agent in framing her moral problems by sensitizing her to the salience of some traits, locating her weaknesses, identifying her interlocutors, focusing the possible objects of attention, and providing the normative vocabulary for her to narrate her own story. These are complex normative tasks, which require ethical theory. That is because only ethical theory addresses the following question: How can we make ourselves better (Murdoch 1999,

[33] Pragmatists and empiricists would agree that the search for elegant, simple, and general theories is driven by practical reasons, by our needs and interests. On this point, see Railton (1991, pp. 187-188).

p. 368)? There are competing answers to this question, and my aim here is not to defend any in particular, but to make the case that it is a question intelligent moral agents should ask. The distinctive purpose of an ethical theory is to provide us with a plausible and morally decent ideal of agency.

We should reconsider in this new perspective how ethical theory contributes to our understanding of moral phenomena. If we are interested in explaining why somebody holds racist beliefs and how racism emerged, we should read psychological and sociological reports on the subject. Williams and Baier are correct that ethical theory contributes nothing to this investigation. More precisely, ethical theory does not address this issue at all: rather, its purpose is to account for the reason why racism is morally objectionable. The kind of contribution that ethical theory provides to our understanding of racism is a moral reconstruction, which identifies the racist's moral failure. If we want to understand what is morally wrong with racism, then there is no other way than referring to ethical theory. Of course, different normative theories represent the racist's moral failure through different (sometimes competing) conceptual articulations. Consider the following moral reconstructions:

(a) the racist commits a sin because God created all equal,
(b) the racist commits a category mistake,
(c) the racist fails to universalize because he does not understand the logic of moral language,
(d) the racist fails to universalize because he lacks the adequate criteria of salience,
(e) the racist fails to universalize because he is lacks a good upbringing,
(f) the racist fails to universalize because his society is such that it inhibits mutual recognition,
(g) the racist fails to universalize because of lack of will,
(h) the racist simply does not care.

These are competing accounts of what goes wrong, morally speaking, with the racist. Each of them predisposes us to see the moral problem in a certain way and invites us to accept a certain solution. For example, by accepting the moral explanation (d), we are encouraged to focus on moral education and identify the solution in terms of the individual's change in habits and rules. If we accept (c), we would consider racism as a kind of logical inconsistency to be exposed by rational argumentation. If we accept (e), we would think of racism as a social problem to be addressed

by modifying the social practices, beliefs, and institutions that promote it.

Similarly, ethical theory addresses conflict as a moral issue through specific kinds of articulations. If one thinks of moral conflicts as generated by value pluralism, one might not want to search for a method of reasoning that systematically undermines the diversity of values. Alternatively, one may concentrate on conflicts such as Antigone's and argue that they arise because of the specific features of the agent's society. In this case, ethical theory would not be invoked to determine whether the duty to obey Creon's law trumps the duty to honor family ties. Rather, ethical theory would help supersede (rather than resolve) the conflict by identifying its social and institutional sources. These two competing philosophical explanations of moral conflict shape competing accounts of the tasks that ethical theory is designed to accomplish. Thus, normative ethical theories importantly differ in the solutions they propose (or do not propose) for moral conflict because they importantly differ in their understanding of moral conflicts. Understanding conflict as a moral phenomenon is a task that ethical theory must address before instructing the agent about whether or how to resolve it.

To be sure, some of these articulations are more useful and plausible than others, but this is not the issue at stake here. My point is that ethical theory is designed to offer full articulations of accounts of this sort. Such articulations could be fairly elaborated. For example, instead of privileging one single account, ethical theory might represent racism as a very complex phenomenon, which requires us to take into account simultaneously the individual's capacities to categorize and universalize, acquire appropriate criteria of salience and develop a strong character, and the features of a society that favor the development of these capacities.[34] Similarly, when addressing the issue of moral conflict, ethical theory might be able to sort out a complex taxonomy of moral conflicts, with a distinctive phenomenology and distinctive normative features. To be useful the articulations provided by ethical theory should rely on other theoretical resources and theories such as sociology, biology, and psychology, as Baier and Williams suggest. However, ethical theory is designed to fulfill a distinctive task, which is that of providing normative standards for judging whether the agent's character, emotions, and actions are morally appropriate.

This is not to say that by providing such normative criteria and supplying the canons of moral reasoning ethical theory thereby succeeds

[34] I therefore reject Annette Baier's suggestion that to undertake a moral ideal is to point at one single source of imperfection; see Baier (1985, p. 214).

in convincing the racist to abandon his racist beliefs, adopt a new attitude and reform his behavior. But it is not obvious that ethical theory addresses the racist who is not troubled by his racism, sensitive to others' criticism and willing to be persuaded by rational argumentation, as it is not obvious that ethical theory addresses the complete amoralist. According to a venerable tradition, for example, arguments are lost on people who are not already perceptive of the difference between a noble and shameful behavior.[35] Without taking such a strong stand, many other ethical theories assume they are speaking to persons endowed with a minimal moral sense. Thus, it could be true, as Williams and Baier maintain, that ethical theory is of no use in reforming the racist, but it is not obvious whether this is what ethical theory was designed to do. It is enough that ethical theory provides a valid account of why racism is a morally objectionable behavior and to those who have already taken an interest in morality it gives reasons to condemn.[36] To deprive these agents of ethical theory is to deplete their moral vocabulary and conceptual resources, and therefore to impoverish their capacity to understand their experience.

One may rebut that the loss of concepts is not always detrimental to our understanding of moral phenomena; in fact, sometimes it is saluted as a mark of moral progress. For example, one may maintain that the concept of "sin" is better lost than used. However, this amounts to saying that some particular moral explanations (such as those that use the concept of sin) are inadequate. Even if such claim were correct, it is no proof that ethical theory does not contribute to our understanding of moral phenomena. To the contrary, its contribution is important and unique.

---

[35] "Arguments seem to have enough influence to stimulate the civilized ones among the young people, and perhaps to make virtue take possession of a well-born character that truly loves what is fine; but they seem unable to stimulate the many towards being fine and good. For the many naturally obey fear, not shame; they avoid what is base because of the penalties, not because it is disgraceful. For since they live by their feelings, they pursue their proper pleasures and the sources of them, and avoid the opposed pains, and have not even a notion of what is fine and [hence] truly pleasant, since they have had no taste of it. What argument could reform people like these?" (Aristotle 1985, 1179b5-17).

[36] This might seem a small consolation, for the immoralist can simply say: Who cares? As Robert Nozick argues, just this lack of care gives already the measure of how much the immoralist pays for his conduct: "The immoral person thinks of getting away with something, he thinks his immoral behavior costs him nothing. But that is not true; he pays the costs of having a less valuable existence. He pays that penalty, although he does not feel it, or he does not care about it. Not all penalties are felt" (1981, p. 409); see also his (1981, pp. 405-406, 410-413, 430).

What I have been suggesting is that understanding oneself and one's experience is a moral task, whose pursuit is furthered in a distinctive way by ethical theory. To renounce ethical theory deprives us of the theoretical resources necessary to address a most fundamental question: how to make ourselves better? It is because such question is so important and central to our lives that ethical theory is part of our common moral practices. In this sense, the kind of justification that ethical theory provides does not requires us to bracket or get out of our practices, but only to stand back from them and be willing to stand scrutiny.[37]

That there is continuity between the theory and the practice of morality should not make us underestimate the challenge that theories posit to who we are and what we do. Undertaking a moral ideal does not leave things as they stand. Ideals challenge our motivations, habits, practices, and convictions, and therefore exert a potentially unsettling influence on us. This is not because morality is a coercive institution designed to preempt or contain anti-social and other sanctioned kinds of conduct. Nor is this because theorizing is the kind of contemplation that inevitably corrodes our convictions and undermines the cohesion and coherence of our practices. Rather, this is because to undertake a moral ideal is to entertain the thought that things should be otherwise, that we might look at things differently, envision alternative courses of action, and adopt other attitudes. This exercise of theorizing might or might not reaffirm the motivations, practices, and convictions we have; and whether it does it or not, it is after all, irrelevant. What matters is that as the outcome of this exercise of theorizing, we have acquired a deeper understanding of ourselves, of the motives that do and should drive us, and of the many sources of our weaknesses and strengths. By offering us a moral ideal, ethical theory refocuses our attention and makes new places for reflection.

To appreciate this power of ethical theory, we must give up the idea that ethical theory is necessarily a monolithic fabrication, the "outcome

[37] This is also to respond to Michael Walzer claim that to conceive of ethical theory as positing this challenge (what he names the "path of invention") is to dismiss or fail to appreciate its actual origins and roots in communal practices, see his (1987, pp. 3-32). The issue of the relation between theory and practice has been the topic of intense methodological discussion, since John Rawls's defense of the method of Reflective Equilibrium, and his seminal essays on the structure of ethical theory, Rawls (1951) and (1974). I cannot enter this vast debate here, but I want to note that my proposal is designed to stress the ideal feature of ethical theory, and insists on the importance that ideals remain ideals.

of contemplating some unified structured world,"[38] the philosopher's flight from reality, ignorant or oblivious of communal practices. Deepening our understanding of ourselves marks a moral change, that is, a change in the way we view things and relate to ourselves and to others. This is not only to say that we actually do experience morality as requiring a critical assessment of our views and relations, but also that theorizing is one important way of practicing and living up morality.[39] This is my answer to Williams' charge that ethical theory necessarily encourages simplification and narrowing one's possibilities. On the contrary, I argue, ethical theory allows us to broaden our views and appreciate alternatives we did not consider at first. As Iris Murdoch eloquently puts it, theorizing helps us 'refresh the tired imagination of our practice', and makes us grasp a wider array of possibilities (Murdoch 1999, pp. 180-181, 184).

5. The Intelligibility of Moral Experience as a Requirement of Adequacy for Ethical Theory

The most distinctive purpose of ethical theory is to present us with moral ideals of agency, which are unsettling as much as they are inspiring. The issue in assessing the viability of ethical theory is whether it posits challenges that it is worthwhile for us to undertake.[40] To serve its purpose well and be of assistance, ethical theory should offer a plausible and decent ideal of moral agency. It is therefore on these two bases that ethical theory must be evaluated: we should ask whether it purports a plausible picture of who we are, and points us toward better selves.

---

[38] The expression is Baier's but it is representative of the view held by many anti-theory philosophers, see her (1985, p. 233).

[39] To this extent, I actually agree with Baier that ethical theories are the myths of philosophers. Myths and theories share this feature of picturing an alternative, and have the effect of modifying the perception of how things stand. In defending the need for ethical theory, there is no point in denying that ethical theories are cultural products. Compare Baier (1986, p. 539).

[40] Iris Murdoch puts it eloquently: "Moral philosophy is the examination of the most important of all human activities, and I think that two things are required of it. The examination should be realistic. Human nature, as opposed to the nature of other hypothetical spiritual beings, has certain discoverable attributes, and these should be suitably considered in any discussion of morality. Secondly, since an ethical system cannot but commend an ideal, it should commend a worthy ideal. Ethics should not be merely an analysis of ordinary mediocre conduct, it should be a hypothesis about good conduct and how this can be achieved. How can we make ourselves better? Is a question moral philosophers should attempt to answer" (1970, p. 78).

Against the background of the view of theorizing I defended, I argue that ethical theory should aim at accounting for the agent's experience in a way that must be intelligible to the agents themselves. The considerations that the agent provides in particular cases matter in order to evaluate the moral relevance, significance, and appropriateness of the agent's experience. For example, in order to assess whether Ingrid's guilty feelings about having sex out of wedlock are meaningful and appropriate, the theorist investigates the kind of account that Ingrid offers of them. In this investigation, Ingrid's considerations about her education, history, and ties to her community might be decisive in offering an intelligent reconstruction of her case.

The philosopher's job in these investigations is to provide an intelligent reconstruction of the agents' moral experience that these agents can conceive as theirs. I will call this the condition of intelligibility. This condition allows me to answer the question as to how the appeal to moral phenomenology provides grounds for criticizing ethical theory: it does so when such a theory violates the condition of intelligibility. The agent's experience constitutes a reason for reviewing or rejecting an ethical theory, a falsifying factor, when the representation that the theory allows for the experience of the agent is not intelligible to the agent herself, and there is no independent ground for considering the agent morally incompetent.

To endorse the requirement of intelligibility is to reaffirm that there is continuity between living morally and theorizing about it. But isn't this criterion exposed to the charge of circularity, as anticipated in section 3? In order to reply to the objection, it is important to recollect the conception of moral phenomenology at work here. Feelings entertain complicated and varied relations with our thoughts and beliefs, but one should not be encouraged to think that their importance is limited to expressing beliefs. Rather, moral feelings are complex intentional activities, which are responsive to reasons and sensitive to judgment. Thus, ethical theory should provide us with normative criteria for understanding and assessing our feelings. The requirement of intelligibility makes sense against the background of a philosophical account that recognizes moral feelings as complex practical activities, (rather than, e.g., emotions merely associated with the judgment). This conception of feelings as complex practical activities allows us to eschew the charge of circularity because it calls attention onto the normative criteria for the assessment of feelings. To be able to guide us, these criteria ought to be based on a plausible psychology; but to guide us well, these criteria ought to represent a decent ideal of moral agency.

Because the nature of these criteria is normative, the intelligent reconstruction that the ethical theorist sets to provide is likely to be revisionist. For any moral agent such reconstructions would be a challenge and a source of inspiration. For example, to Matilda, the rape victim tortured by guilty feelings and self-blame, ethical theory might provide an alternative normative understanding of the case, which helps her to redirect her attention to having being wronged, and consequently relocate her blame on the perpetrator. To adopt an appropriate reaction to the situation, Matilda is required to develop an intelligent account of her predicament, through which to identify the proper object of her reaction. To direct one's feeling toward the appropriate object is a moral exercise and, in Matilda's case, it requires a revised account of the problem. To let emerge and endorse the appropriate kind of emotion, such as, for example, anger or grief, is also a moral achievement, which in Matilda's case requires a revised narrative, but it does not require that she reclassify her experience. Thus, this revision does not amount to discounting Matilda's moral experience, but to providing her with a normative structure to intelligently understand it.[41] Accepting this revision marks a moral change for Matilda, that is, a change in self-awareness.

In other cases, ethical theory encourages the agent to discern and unfold the multiplicity of relations between her emotional experience and her judgment about the case. For example, suppose Luce judges her guilty feelings about not attending the departmental meeting as misplaced and almost obsessive. By suggesting that feelings are typical modes of self-assessment, ethical theory might offer a different reconstruction of the significance Luce's experience. Against the background of these new theoretical considerations, Luce might come to realize how much she values and identifies with her profession, and become more aware of her priorities. On this occasion, theorizing helps the agent to give more credit to her own experience. Like in Matilda's example, to endorse the revision marks a moral change, and more precisely, a deeper level of self-awareness.

These examples of revisionist theoretical reconstructions of our moral experience show that to insist on the condition of intelligibility as a requirement of adequacy for ethical theory is not to claim that the agent is beyond criticism, or to admit that the criteria for assessing the agent's

[41] This responds to Baier's complaint that convictions and experience of the agent "serve as a very weak constraint on contemporary theory, something to take note of, then either give reasons to endorse or explain away as understandable but misguided judgment" (1986, p. 539).

experience could only be internal. These criteria of appropriateness are normative, and represent ideal standards for judging, even when we subject them to the condition of intelligibility. Reflective agents are capable of assessing their own experience, but they need theoretical guidance in mapping their own alternatives. Theorizing stretches their imagination and makes them consider new options, far beyond the limits that she used to take as ultimate.

The claim that we do not need anything like an ethical theory, although we need other theoretical considerations, is therefore based on a misunderstanding of the purposes of theorizing. The search for understanding is a moral search, to be guided by an ideal of moral agency. Ethical theory cannot do more than offering guidelines as to how to intelligently reconstruct our moral experience and learn from it. Instead of narrowing down our options, as Williams suggests, ethical theory expands them by challenging the bounds set by our habits and convictions. It makes us reflect on how we could be better, given (not independently of) who we are.

It is to assist the agents in their search for self-understanding that ethical theory provides elucidations, analyses, and normative reconstructions. This is a modest yet noble task for an ethical theorist to undertake. My purpose in this paper has been to reclaim it.

Acknowledgements

A distant ancestor of this paper was presented at the University of St. Andrews and at Stirling University in September 1999. I would like to thank my audiences and, in particular, John Broome, John Cullity, John Skorupski, and Suzanne Uniacke. I owe thanks to Elijah Millgram, Amélie Rorty, Robert Schwartz, and Sergio Tenenbaum, for their precious comments on previous drafts. Further work has been possible thanks to a fellowship at the Center for Twentieth Century Studies at the University of Wisconsin Milwaukee in 1999 and a research grant in 2000, a Research Award of the University of Wisconsin Foundation and Graduate School, and a fellowship at Institute for Research in the Humanities at the University of Wisconsin at Madison in 2002-2003.

University of Wisconsin-Milwaukee
Department of Philosophy
3243 N Downer Ave
Milwaukee, WI 53211, USA
e-mail: cbagnoli@uwm.edu

REFERENCES

Aristotle, (1985). *Nicomachean Ethics*. Indianapolis: Hackett Publishers.

Bagnoli, C. (2003). Respect and Loving Attention. *Canadian Journal of Philosophy* **33**, 483-516.

Baier, A. (1985). Doing Without a Moral Theory? In: *Postures of the Mind: Essays on Mind and Morals*, pp. 228-247. Minneapolis: University of Minnesota Press.

Baier, A. (1986). Extending the Limits of Moral Theory. *The Journal of Philosophy* **83**, 538-545.

Barcan Marcus, R. (1980). Moral Dilemmas and Consistency. *The Journal of Philosophy* **77**, 121-136.

Dancy, J. (1985). The Role of Imaginary Cases in Ethics. *Pacific Philosophy Quarterly* **66**, 141-153.

Donagan, A. (1996). Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy. In: Mason (1996), pp. 11-23.

Fingarette, H. (1979). Feeling Guilty. *American Philosophical Quarterly* **16**, 159-164.

Foot, P. (1995). Moral Dilemmas Revisited. In: Sinnott-Armstrong *et al.* (1995), pp. 117-128.

Foot, P. (2002). *Moral Dilemmas*. Oxford: Oxford University Press.

Gibbard, A. (1990). *Wise Choices, Apt Feelings.* Oxford: Clarendon.

Gibbard, A. (1995). Why Theorize How to Live with Each Other? *Philosophy and Phenomenological Research* **55**, 323-342.

Gowans, C.W. (1996). Moral Theory, Moral Dilemmas, and Moral Responsibility. In: Mason (1996), pp. 199-206.

Greenspan, P.S. (1994). *Practical Guilt*. Oxford: Oxford University Press.

Greenspan, P.S. (1995). Perspectival Guilt. In: Sinnott-Armstrong *et al.* (1995), pp. 46-59.

Hare, R.M. (1981). *Moral Thinking*. Oxford: Oxford University Press.

Herman, B. (1993). *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.

Hill, T.E. (1996). Moral Dilemmas, Gaps, and Residues: A Kantian Perspective. In: Mason (1996), 167-198.

Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

Mason, H.E., ed. (1996). *Moral Dilemmas and Moral Theory*. Oxford: Oxford University Press.

McConnell, T. (1996). Moral Residue and Dilemmas. In: Mason (1996), pp. 36-47.

McIntyre, A. (1990). Moral Dilemmas. *Philosophy and Phenomenological Research* **50**, 367-382.

Morris, M. (1992). Moral Conflicts and Ordinary Emotional Experience. *Journal of Value Inquiry* **26**, 223-227.

Mothersill, M. (1996). The Moral Dilemmas Debate. In: Mason (1996), pp. 66-85.

Murdoch, I. (1970). *The Sovereignty of Good*. London: Routledge.
Murdoch, I. (1999). *Existentialists and Mystics*. London: Penguin Books.
Noble, C. (1989). Normative Ethical Theories. In: S. Clarke and E. Simpson (eds.), *Anti-Theory in Ethics and Moral Conservativism*, pp. 49-64. New York: Albany Press.
Nozick, R. (1981). *Philosophical Explanations.* Cambridge, MA: Harvard University Press.
Railton, P. (1991). Moral Theory as a Moral Practice. *Nous* **25**, 185-190.
Railton, P. (1992). Pluralism, Determinacy and Dilemma. *Ethics* **102**, 720-743.
Rawls, J. (1951). Outline of a Decision Procedure in Ethics. *Philosophical Review* **60**, 177-197.
Rawls, J. (1974). The Independence of Moral Theory. *Proceedings of the Addresses of the American Philosophical Association* **47**, 5-22.
Ross, D.W. (1930). *The Right and the Good*. Oxford: Oxford University Press.
Scanlon, T. (1992). The Aims and the Authority of Moral Theory. *Oxford Journal of Legal Studies* **12**, 1-23.
Scanlon, T. (1995). Moral Theory: Understanding and Disagreement. *Philosophy and Phenomenological Research* **55**, 343-356.
Sinnott-Armstrong, W. (1988). *Moral Dilemmas*. Oxford: Blackwell.
Sinnott-Armstrong, W., D. Raffman and N. Asher, eds. (1995). *Modality, Morality, and Belief: Essays in Honor of Ruth Barcan Marcus*. New York: Cambridge University Press.
Statman, D. (1995). *Moral Dilemmas*. Amsterdam: Rodopi.
Van Fraassen, B. (1973). Values and Heart's Commands. *The Journal of Philosophy* **70**, 5-19.
Walzer, M. (1987). *Interpretation and Social Criticism*. Cambridge, MA: Harvard University Press.
Williams, B. (1963). Ethical Consistency. *Proceeding of the Aristotelian Society* (supplement) **39**, 103-124. Reprinted in: Williams (1973b), pp. 166-186.
Williams, B. (1965). Consistency and Realism. *Proceeding of the Aristotelian Society* (supplement) **60**, 1-22. Reprinted in: Williams (1973b), pp. 187-205.
Williams, B. (1973a). Consistency and Realism. Additional Note. In: Williams (1973b), pp. 205-206.
Williams, B. (1973b). *Problem of the Self: Philosophical Papers 1956-1972*. Cambridge: Cambridge University Press.
Williams, B. (1981). *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press.
Williams, B. (1985). *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Philip Clark

HOW REASON CAN BE PRACTICAL: A REPLY TO HUME

Beliefs can only work with the goals we have. Take the belief that there is a cat in the kitchen. This belief might move you toward the kitchen, or away from it; but wherever you wind up, it won't be the belief alone that gets you there. The belief will work with some "designed end or purpose" (Hume 1888, p. 414), like petting the cat, or avoiding it. And so it is with all beliefs, or so I take Hume to say.

Now I doubt very much that this doctrine would ever have been questioned if it weren't for certain assumptions about where it leads. Hume was happy enough with it, because he didn't mind denying that reason could combat a passion. Others have been happy enough with it, because they weren't wedded to the idea of objectivity in ethics; others because they weren't wedded to the idea that value judgments have a necessary grip on the will. But I do think reason can combat a passion, and I am wedded to the idea of objectivity in ethics, and I do think value judgments have a necessary grip on the will. Still, I do not feel like questioning what Hume said about beliefs. It seems to me that Hume's famous doctrine doesn't lead much of anywhere.

The reason it doesn't is that it leaves us a great deal of latitude in constructing a theory of practical reasoning. As long as we conceive of practical reasoning as an activity in which the reasoner pursues a goal, we can accommodate the idea that beliefs can only work with the goals we have. But the idea that practical reasoning is a goal-directed activity does little to restrict our options in ethics. It leaves plenty of room for a view on which reason combats passions, values are objective, and value judgments entail a disposition of the will. The trick is to occupy this room. For a satisfying answer to Hume, I think we have to stop trying to explain how beliefs can motivate by themselves, and start trying to explain how it doesn't matter whether they can or not. This puts me at

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 213-230. Amsterdam/New York, NY: Rodopi, 2007.

odds with Nagel (1970), McDowell (1978), Platts (1979, 1981), Dancy (1993), McNaughton (1988), and Little (1997), all of whom try, in one way or another, to explain how beliefs can work alone. It aligns me more with David Velleman (1992a, 1996), when he speaks of agency as having a constitutive goal, and with Gavin Lawrence (1995), when he says practical reasoning has a formal object.

Suppose, then, that we are going to use the idea that reasoning has a goal to explain the motivational properties of ethical thought. What will we need to say about reasoning, if we want to say reason can combat passion, values are objective, and value judgments entail a disposition of the will? My thesis here is not that we should say all these things. It is that we can say all these things, even if we keep our fingers off of Hume's doctrine about belief.

Truth is one thing at which reason is often said to aim. When I reason about why the dinosaurs became extinct, say, my aim is to follow reasons to a true conclusion. If all goes well, I come away with a true belief about why the dinosaurs became extinct. Ethics does not appear to be immune to this kind of reasoning. People do sometimes try, at least, to follow reasons to true conclusions about what they ought to do, or what would be best. But the goal of truth isn't going to help us explain why "ought" judgments grip the will. It will explain, at most, why we arrive at the "ought" judgments we do. We will have to tell a separate story about where they get their special influence on the will.

One might object, at this point, that I am simply assuming that the world is "inert," in other words, that there are no features of the world whose recognition entails that the agent is disposed to act in one way rather than another. That would be unfair to those who think of the world as "ert." But I am not assuming that the world is inert. Rather, I am assuming, at least for the sake of argument, that the world is ert. Suppose knowledge of how we should act does entail a disposition of the will. My question is how we might use the idea of a goal of reason to explain why that is so. The goal of truth, I am saying, is not going to help, because it can only explain why we arrive at the ethical judgments we do. It can't explain why those judgments have motivational magnetism.

But there is another way in which reason is sometimes said to bear on action. We sometimes choose an action for its connection with something we want. If we were thinking out loud on these occasions, we might say something like, "I can spare Bob's feelings by saying I like his singing, so I'll say that," or "I can't honestly say I like Bob's singing, so I won't." Here the "so" signals the formation of an intention, grounded in a belief connecting the action with something the agent wants. Could this picture

of practical reasoning help explain the motivational magnetism (Stevenson 1937) of ethical judgments? As it stands, it is hard to see how. But as it stands the picture is incomplete. We haven't yet seen how ethical judgments would fit into such a scheme. So far we see the role of beliefs like "I can spare Bob's feelings by saying I like his singing," and "I can't honestly say I like Bob's singing." But we are out to explain the motivational properties of judgments like "It is important not to hurt people's feelings," and "Truthfulness is something I have reason to pursue." These judgments come in when we add an account of deliberation.

Almost any occasion for reasoning about what to do will put the agent in the position of deciding which arguments to act on, and which to let slide. Take Eleanor, who has just been asked point blank, by Bob, what she thinks of Bob's singing (Clark 2001a, pp. 494-500). Eleanor might try to spare Bob's feelings, by saying she likes Bob's singing. And she might decide to risk hurting Bob's feelings. That would be the honest path, since she is in fact appalled by Bob's singing. Either way, she will be choosing a course of action for its connection with something she wants. She could resolve the issue by flipping a coin. But another option would be to deliberate. That is, she could weigh the arguments for and against the available options, in an effort to follow the weight of reasons.

Now deliberation, so construed, is a reasoning activity in which one pursues a goal. The goal is, roughly, to follow the weight of reasons. This looks like a goal that could do the work we have in mind. A judgment like "Truthfulness is something I have reason to pursue" lies directly in its path. Perhaps the magnetism of that judgment, as compared with, say, the belief that there is a cat in the kitchen, can be explained by reference the goal of deliberation.

But there is an immediate difficulty. It looks as if we are saying that practical reasoning both is and is not aimed at discovering the truth about what one ought to do. When Eleanor deliberates, she will reason about whether she ought to spare Bob's feelings, or whether it might be better to tell the truth. In this reasoning, she will aim for a true conclusion about what she ought to do, or what would be best. But we are also thinking that in the end she will choose a course of action for its connection with something she wants. She will go, "By doing this I can do that, so I'll do this." And that doesn't look at all like an attempt to reach a true conclusion. It looks like what Hume calls choosing a means to an end.

It isn't at all clear how these two phenomena are supposed to merge in a reasoning activity whose goal could explain the magnetism of ethical

thought. They look like distinct activities. Each, taken separately, does appear to have a goal of its own. On the one hand, reasoning about what is best aims at the truth about what is best. On the other hand, when you choose a course of action for its connection with something you want, you aim to align your behavior with the thing you want. Not every action is equally well suited to a desired end. The idea is to pick one that suits. The problem is that neither of these goals is going to do the work I want it to do. I want to be a hard-core rationalist objectivist internalist. And I want a goal of reasoning to explain the internalism. But the goal of truth isn't going to explain the internalism at all. And the goal of fitting my behavior to my desires, while it might explain the internalism, will thereby cost me my objectivism. Allow me to illustrate.

So far we have on the table two models of deliberation. On one model, deliberation aims for a conclusion about what one ought to do. The competing arguments weighed in deliberation are thus weighed for their relevance to the topic of what one ought to do. For instance, the fact that Eleanor can't honestly say she likes Bob's singing will be weighed as a reason (prima facie or pro tanto) for thinking Eleanor ought to hold her tongue. The whole activity thus funnels down to a grand conclusion about what one ought to do. In this kind of reasoning, Eleanor will aim to follow reasons to a true conclusion about what she ought to do. But that aim can't help us explain the motivational force of the resulting "ought" judgments. If there is a goal of reason that can explain why "ought" judgments dispose us to action, the goal of believing the true is not it.

On the other model, what the deliberator does is choose a course of action for its connection with something she wants, or some set of things she wants. Suppose Eleanor wants to spare Bob's feelings, but isn't keen on deliberately misrepresenting her attitude. In deliberation she looks to other desires that might settle the matter. For instance, she might want to spare Bob future humiliation. If so, she might reason that a little humiliation now could spare him a load of humiliation later. And that might lead her to choose truth telling, for its connection with honesty and Bob's future dignity, and in spite of its connection with Bob's mood in the short run. On this model, Eleanor aims in her deliberation to establish a certain sort of fit between her behavior and her desires.

Now how could we harness that aim, to explain the motivational magnetism of ethical judgments? Suppose Eleanor thinks it is important to be truthful. We might construe this as a thought about the fit between Eleanor's desires and truthful behavior on Eleanor's part. That is, we might say that what Eleanor is really thinking is that truthful behavior fits her desires. Given that the goal of deliberation is to fit one's behavior to

one's desires, this thought is uniquely positioned to work with that goal. We can say that insofar as Eleanor deliberates at all she will pursue the goal that would join with this belief to issue in action. That is not something we can say about just any belief. So we have our explanation of how the judgment gets its special purchase on the will. Alternatively, we might deny that the thought is really a belief. We might say it just is a desire. That too would explain how the judgment automatically engages the goal of deliberation.

But of course I can't say these things. They offend against my objectivism. For me, the thought that truthfulness is important is a genuine belief. Moreover, it is not a thought about one's own desires. The importance of honesty does not rise and fall as my desires gather and disperse behind honest behavior. Rather, honesty is a standard against which my desires can be judged. But for precisely this reason, the goal of fitting one's behavior to one's desires cannot explain the magnetism of the thought that honesty is important. A person might take an interest in fitting her desires to some objective standard, but on the present model there is nothing in the nature of deliberation that requires it.

So I am still looking for a way of understanding what we are doing, when we reason about what to do, on which the object of that reasoning is properly positioned to explain the motivational efficacy of ethical beliefs. Here is a proposal.

Let us think of "ought" judgments, or judgments about what is best, as conclusions about conclusions. On this view, the thought that you ought to say you like Bob's singing is the thought that the weight of reasons is behind the conclusion "So I'll say it." And the thought that it would be better not to is the thought that the weight of reasons is behind the conclusion "So I won't." The question where the weight of reasons lies is a topic on which one can reason, and in reasoning about that one will aim for a true conclusion about where the weight of reasons lies. But reaching a true conclusion about where the weight of reasons lies is one thing, and giving the reasons the weight they deserve is another. The distinctively practical goal that we need to harness, I suggest, is the goal of giving the reasons the weight they deserve. Deliberation should be conceived as an activity in which one weighs arguments for and against different courses of action, in an attempt to follow the weight of reasons to a conclusion of the form "So I'll do such and such."

When Eleanor deliberates about how to answer Bob, on this view, she considers various steps she could make (Clark 1997). She could go "I can spare Bob's feelings by praising his singing, so I'll do that," and she could go "I can't honestly say I like Bob's singing, so I won't." Later on,

it may occur to her that she could go "For one thing, I can't honestly say I like Bob's singing, and for another, a little embarrassment now could save Bob a load of humiliation later, so I'll tell him what I think." Her aim in deliberation is to give each of these possible steps, and any others that may come into view, the weight it deserves. In the end she may reach the conclusion that a certain step deserves to win. That is, she may decide that deliberating well in the circumstances means making that step. This is a conclusion about the weight of reasons. And now, since she has the aim of following the weight of reasons, she has a goal and a belief that can join to explain why she makes the step she does. If she decides, for example, that considerations of honesty deserve to win, then since she is out to make the steps that deserve to win, she has a goal that can explain why she makes the step from honesty, rather than the step from kindness.

The idea that at least some ethical beliefs are conclusions about conclusions has a twentieth-century precedent in the work of Donald Davidson (1980; see also Clark 2001a, p. 495). On Davidson's view, deliberation reaches a judgment about what conclusion has the support of the available evidence, taken as a whole. For instance one might reach the conclusion that one ought, in view of the available evidence, to draw – or "detach" – the conclusion that it would be best to quit smoking. Weakness of will happens, on this view, when one reaches the "all things considered judgment" that one ought to detach the conclusion that it would be best to quit smoking, but fails to detach the "unconditional judgment" that it would be best to quit smoking. Our view follows this pattern, with one significant adjustment. For Davidson, the all things considered conclusion and the unconditional conclusion are both judgments about what is best. So if we are to follow Davidson, we must believe that there are two distinct kinds of judgment about what is best. On the view I have in mind, there is just one sort of judgment about what is best, and that is the all things considered judgment – the judgment about where the weight of reasons lies. That is a judgment about what steps to make. But the steps themselves do not conclude in beliefs about what is best. They conclude in the formation of intentions. We are not saying, as Davidson is, that there are two kinds of judgment about what is best.

So here is a goal, the goal of making the steps that deserve to win, that is different from either of the goals we've considered so far. The goal of making the steps that are warranted is not identical with the goal of reaching a true conclusion about which steps are warranted. Information about which steps are warranted is deadly relevant to the project of making the warranted steps. But the project is not simply to

gather the information. So one's goal in deliberation is not simply to reach a true conclusion about where the weight of reasons lies.

Likewise the goal of making the steps that deserve to win is not identical with the goal of fitting one's behavior to one's desires. There may be truths about which steps are warranted that cannot be read off of my desires. If so, the goal of deliberation will require me to look beyond my subjective motivational set (Williams 1970). It is true that we can make the goals the same by adopting a subjectivist analysis of truths about which steps are warranted. If we assume that these truths can be read off my own desires, then the goal of deliberation will not require me to look beyond my desires. But we objectivists will think it does.

How might this goal help explain the motivational magnetism of ethical thought? Well, suppose Eleanor thinks of honesty as a consideration that deserves weight in the circumstances. She thinks, to put it roughly, that the step from honesty has a defeasible claim to be the step that deserves to win. We want to explain how this belief can have a necessary grip on the will. The explanation is that what Eleanor believes connects up with the goal of deliberation. The same cannot be said for other beliefs. If Eleanor believes there is a cat in the kitchen, very little follows about how that belief is going to play in her thinking about what to do. But if Eleanor believes honesty deserves weight, then we can say that insofar as she deliberates about how to answer Bob, her goal will be one that engages the belief. She will be out to make the steps that deserve to win, and she will regard the step from honesty as having a defeasible claim to be the step that deserves to win. This explains the special grip of judgments about what deserves weight in deliberation.

Notice that the explanation does not say the judgments can work by themselves. It says the judgments work by joining up with a goal that drives practical reasoning. So we are not envisioning beliefs that somehow generate motivation out of thin air. We are granting Hume his point that beliefs can only work with the goals we have.

Now I promised a view that would grant Hume his point about beliefs, without compromising the objectivity of ethics, without playing down the motivational magnetism of ethical judgments, and without denying that reason can combat a passion. Let me speak briefly to each of these points.

First, one might suppose that our appeal to a goal scotches the objectivity of ethics, because the truth of ethical claims must now rest on the fact that the agent has the goal. One might suppose, in other words, that the values we have in mind must now be hypothetical imperatives, or internal reasons. That would be a mistake. On the present picture, the

importance of truthfulness does not require that Eleanor have the goal that is pursued in deliberation. Eleanor might just blurt something out, without making any effort to follow the weight of reasons. The reasons can still be there, because there can still be truths about what good deliberation would look like. Eleanor doesn't have to go in for deliberation, in order for these truths to exist.

Second, one might think we are now committed to a watered down version of internalism. I'm not sure exactly what should count as strong enough, but I propose the following standard. If it turns out that having an ethical belief is comparable to having a belief-desire pair, in terms of what follows about how one will act, then we have a strong enough version of internalism. So we'll take belief-desire pairs as the paradigm case of an essentially motivational state, and we'll ask whether we can explain how having an ethical belief could be as good as having a belief-desire pair, in terms of what follows about how one will act (Clark 2000, pp. 375-378). First, consider what might be said about the motivational implications of belief-desire pairs.

If I want to hear a loon, and believe that I can do this by holding very still, and only by holding very still, then I will choose to hold very still, unless the mosquitoes are unbearable, or I have a seizure, or I am distracted by anger, or I am in a hurry and can't take the time, or any of a million other things. One safe thing to say is that I will choose to hold still unless I don't. But some philosophers have thought they could eke out a bit more. They would say that I am in a state that *disposes* me to choose to hold still, or that *other things equal* I will choose to hold still. This is to say, roughly, that belief-desire pairs have a certain motivational spin, or default setting, that will carry the day unless something overwhelms it, or cancels it, or reverses it, or what have you.

On our view, ethical beliefs do have motivational spin. Here is what the possession of an ethical belief entails. It entails that insofar as the agent deliberates she will have the goal of giving the various steps the weight they deserve. This tells us how her deliberations are going to go, other things equal. Since ethical beliefs are about what steps deserve what sort of weight, she is going to make the steps that follow the ethical belief. That is the default setting.

To be sure, there is no guarantee that our agent will deliberate. She might be asleep, for instance. But then, neither will a belief-desire pair have its normal effect in a person who is asleep. If I want to see the New Year's ball drop in Times Square, and know that I can do this by glancing at any working TV at midnight, this will not help me when I fall asleep at 11:55. Similarly, one might engage in practical reasoning but

fail to address the issue to which a particular ethical judgment is relevant. You knew you had reason to wear goggles when cutting glass, but when the time came all you could think about was how to get the shape you wanted. You just started in, as we say, "without thinking." But again, belief-desire pairs can fail in the same way. You wanted to avoid sexually transmitted diseases, say, and you knew how, but at the time you just didn't think.

So although reason judgments can fail to influence choice, due to a failure to use one's practical reason, this does not differentiate them from belief-desire pairs in terms of their *ceteris paribus* implications for action. On our view ethical beliefs, like belief-desire pairs, will carry entailments about how the agent's practical thinking will go, other things equal. And as far as I can see, the other things that have to be equal will be pretty much the same. They will include the use of one's reason. And this means the goal of practical reasoning will ride in on the *ceteris paribus* clause. That is how having an ethical belief can be as good as having a belief-desire pair.

Finally, a word about the combat of reason and passion. Hume looks at reason and sees a search for truth. He looks at action, and sees something that cannot be true or false. This means the goal of reason is not capable of pressing all the way home to action. The search for true conclusions is going to stall out in the formation of beliefs. The search can't reach to actions because actions can't be true conclusions. But if the search for truth leads to a belief, and that belief then leads to action, there has to be a goal that presses home to action, because beliefs can only work with the goals we have. Whatever that goal is, Hume thinks, it won't be a goal of reason. It will be some extra-rational desire. So when you feel like eating a whole box of cookies, there is no goal of reason that could press home to a contrary choice. Reason does have a goal, but that goal can't reach action, so something else will have to do the work.

On our view, though, there is a goal of reason that can press home to action. Sometimes we weigh reasons for and against different courses of action, in an attempt to make the steps that deserve to win. That goal can reach beyond belief to action, because the steps conclude in the formation of intentions. If you feel like eating a whole box of cookies, you can deploy your capacity to reason about what to do. In so doing you will be pursuing a goal that can press home to a contrary choice, by leading you to make a step concluding in "So I won't," or "So I'll go for a swim instead." Neither you nor your reason is a slave to your passions.

So here we have a view that grants Hume's doctrine about belief, but leaves room for a staunch objectivism about ethical truths, explains why

ethical beliefs would entail a disposition of the will, and makes good sense of the idea that reason can prevail over a passion. That means the rationalist objectivist internalists among us should stop taking it out on Hume's doctrine about belief. Likewise, those who accept Hume's doctrine should stop taking it out on us rationalist objectivist internalists. The doctrine just does not bind.

Let it be agreed that a view of the sort I've been describing is available. Even so, we need to see if it can be worked out in a way that makes it philosophically satisfying. On this score, the most obvious puzzler is the talk of truths about which steps deserve weight, or deserve to win. What are these truths? Or is there nothing more we can say about them? In the case of ordinary truth-oriented reasoning, we can say the reasoner aims to follow reasons; but then we can go one better, and say the ultimate aim is to discern the truth. On this picture, reasons for belief are reasons in virtue of their relation to truth. When it comes to reasons for action, can we go one better? Can we say reasons for action are reasons in virtue of their relation to some more ultimate aim of practical reasoning?

One option is to say the object of practical reasoning is to choose well, or to do the best thing. This gives a neat parallel. When we think about what is the case, we aim to follow reasons to a true conclusion. When we think about what to do, we aim to follow reasons to a good choice. If someone wants to know what we mean by "good choice," we refer her to *The Nicomachean Ethics* (Aristotle 1993, Book I). There she will read that choosing well is choosing in a way that is consonant with living well and faring well as a human being. Living well and faring well, reading on, is a matter of having and exercising the excellences of character, in a life that is long enough, where your range of reasonable choices isn't too impoverished (by things like blindness or ugliness or lack of money), and on the whole your choices aren't thwarted by bad luck (like breaking your neck while practicing for the Olympic Games). Now what could be wrong with a view like that?

Well two things, if I read David Velleman correctly (1996). He gives two reasons for thinking the neat parallel isn't so neat. First, there is the fact that agents don't necessarily aim at the good in their actions. Velleman thinks believers do necessarily aim at the truth in their beliefs. So the object of truth-oriented reasoning is something at which believers necessarily aim. But actions are not "necessarily well intentioned" (1996, p. 716). It is possible, and here he follows Stocker (1979), to desire something seen as bad, simply because it is bad, particularly in a mood of silliness, self-destructiveness, or despair. In Milton's *Paradise Lost*, for

instance, Satan reacts to defeat by dedicating himself to the pursuit of evil. In acting from such motives, Velleman says, one does not aim at anything conceived as good. Velleman's worry is that if we stake practical reasoning to the goal of choosing well, rather than to autonomy, we won't be able to do justice to the perversity of Satan's motives. We will wind up portraying Satan as a lover of the good.

The second worry is that "choosing well" fails to specify a substantive object for practical reasoning. Truth looks like a goal that could serve as an independent test of a piece of reasoning. The reasoning will be good if it is the sort of reasoning that is apt to lead to a true conclusion. But talk of "the best action" looks like shorthand for "the action to which good reasoning would lead." What do we mean by "the best action" if not the action best supported by reasons? And what do we mean by "reason," if not something that will be given weight insofar as the agent deliberates well? It is certainly true that good practical reasoning is reasoning that arrives at the action to which good practical reasoning will lead. But if that is all we are saying, when we offer the good as the goal of practical reasoning, then we've failed to offer a test of good reasoning.

Velleman's way with these points is to keep the parallel absolutely neat. He hangs on to the idea that the goal of practical reasoning is something necessarily pursued in action. And he looks for something that, like truth, can provide an independent test of good reasoning. But I want to pursue another way out. I think we should just admit that the parallel is not absolutely neat. The goal of practical reasoning is not something agents necessarily pursue, just in virtue of acting. And the search for a test of good practical reasoning is not going to yield anything so tidy as truth, although it might yield something like Aristotle's outline sketch, paraphrased above. I'll take these points in turn.

How can we accommodate the fact that people don't necessarily aim, in their actions, to do what is best? Note first that activities sometimes inherit the goals of larger activities to which they belong. Consider the activity of making laws. One might hold that in a well-ordered government, legislators make laws with a view to the good of the society. On such a theory, the good of society serves as a standard by which individual acts of legislation can be judged. This is not to say, however, that the good of society is something legislators necessarily pursue, insofar as they legislate at all. In a well functioning government they will have this aim; but in another sort of government they may make laws solely with a view to the good of the tobacco industry. To set up some terminology, we can say the good of society is a *criterion of normative assessment* for law making, but is not a *constitutive aim* of law making.

A criterion of normative assessment is a standard for judging whether a thing is done well. A constitutive aim is a goal one must pursue to be doing the thing at all.

On the view I want to defend, living well and faring well is a goal pursued in good practical reasoning, in roughly the way the good of society is a goal pursued in good government. This goal serves as the standard of normative assessment for actions, but it does not serve as a constitutive aim of action. An agent whose practical reason is functioning properly will aim to act well; that is, to act in ways that have a place in a good human life. But another sort of agent may trade in the aim of acting well for that of having an affair, say, or even for the aim of acting badly. Such actions can be judged against a goal that is not pursued in the actions themselves, but is pursued in good practical reasoning.

Plainly, this view does not commit us to the idea that action is necessarily well intentioned. Satan might act without aiming to act well, just as someone might legislate without aiming at the good of society. In both cases, the performance can be judged against the goal of a larger activity, in one case deliberation, in the other government.

It might seem, though, that if we deny the tight connection between action and doing what's best, we cannot go on saying that practical reasoning is an activity in which one aims to follow the weight of reasons. For if one can act without aiming to act well, then surely one can also reason without aiming to act well. Having chosen some action he knows to be bad, like corrupting some good person's soul, Satan can surely reason about how to do that bad thing. Likewise, having chosen against one's better judgment to have an affair, one can still reason, sometimes with great ingenuity, about how to have the affair. This kind of reasoning goes on in full knowledge that one will not be doing what is best. But if not all practical reasoning aims at doing what is best, then how are we to understand practical reasoning as an activity in which one aims to get it right?

We can see how by getting clearer on what this thought about getting it right comes to. The root idea, I think, is quite general. It applies to ordinary truth-oriented reasoning as well as to practical reasoning. The thought is that weighing reasons is an activity in which one aims to weigh those reasons correctly. In a given piece of reasoning, one aims for that reasoning to go well. But when we go to apply this root idea to practical reasoning, we find that episodes of practical reasoning differ in the kinds of reasons that are being weighed. Sometimes we weigh reasons for and against doing something, full stop, and sometimes we weigh

reasons for and against doing something in one way rather than another. In the latter case one need not have the aim of doing what is best.

To see what I have in mind, consider the akratic smoker, who knows it would be better not to smoke, but lights up anyway. Were she to engage in the activity of weighing reasons for and against lighting up, she would be aiming to weigh those reasons correctly. In that case she would have the aim of choosing the better course of action whichever it might be. But of course, as she positions the cigarette in her mouth and rummages about for her lighter, she need not be deliberating about whether to light up. By then she will ordinarily have decided, against her better judgment, to have another cigarette. So she will have left the goal of doing what is best in the dust, and settled on a goal she herself regards as badly chosen.

But even if she is no longer reasoning about *whether* to have a cigarette, she may reason about *how*. She might consider whether to do it inside or out, for instance; or whether to do it now or wait until the children are in bed. In reasoning about how to go about some action one knows to be bad, one does weigh reasons. One weighs reasons for doing the badly chosen action in one way rather than another. In that reasoning one aims to weigh *those* reasons correctly. But the aim of weighing those reasons correctly is not the aim of doing what is best. It is the aim of choosing the better of two (or more) ways of doing something it would be better not to be doing in the first place. A person who has already abandoned the path she takes to be best can still have this more modest aim.

The sense in which practical reasoning is aimed at getting it right, then, is that in weighing reasons one aims for that very reasoning to go well. We can say that without saying all practical reasoning is aimed at doing what is best.

I turn now to the second problem. When we say practical reasoning aims at a good choice, or the best thing to do, are we in fact specifying a substantive aim for practical reasoning? Or are we just saying practical reasoning aims at what it does? The answer, I think, is that while we are not specifying a useful criterion of normative assessment, we are specifying a substantive constitutive aim.

When we say an activity has a goal, we may be offering the goal as a standard for judging the activity, and we may be offering it as something necessarily pursued in the activity. These are distinct ideas. For instance, Aristotle appears to hold that spear making aims at victory (Aristotle 1993, Book I, especially the remarks on military science at 1094a9*ff*; see also Broadie 1991, p. 12). What he means is that the standard by which to

judge the quality of the work is whether the spears are well suited to the goal of victory. What he *does not* mean is that one cannot make spears without aiming at victory. Aristotle would be the first to acknowledge the possibility of deliberately making a bad spear. A pacifist working in a spear factory might do exactly that, as an act of sabotage. We encountered the same point a while back, when we noted that legislation can be judged against a goal not pursued by the legislators themselves.

So we need to distinguish two things that might be meant by saying practical reasoning aims at the best thing to do. We might be offering the best action as a criterion of normative assessment, and we might be offering it as a constitutive aim. And now we can see that even if the best action fails miserably as a criterion of normative assessment, it can still serve as a substantive constitutive aim. Here is an example to illustrate the point.

Suppose someone puts forward the claim that friendship has being a good friend as a constitutive aim. The thought would be that being someone's friend requires, logically requires, having the goal of being a good friend to that person. (I don't mean the goal of being an especially close friend, but rather the goal of conducting oneself well in one's role as friend – of functioning well as a friend.) I don't know whether this claim is true; but it does seem more plausible for friendship than for, say, being someone's father. Plainly, there are fathers who don't aim to be good fathers. But can we count among our friends people who aren't out to be good friends to us? It's debatable. One thing is certain, however: we can regard being a good friend as a constitutive aim of friendship, without touting it as the criterion of normative assessment for friendship. Such touting would indeed fail to specify any criterion at all. It tells us nothing about what goes into being a good friend. But for all that, there is such a thing as being a good friend. To specify that as a constitutive aim of friendship is to specify a substantive target that must be pursued by anyone who qualifies as a friend at all.

Let us go back, now, to the claim that "practical reasoning aims at figuring out the best thing to do." As we saw, this can be read either as offering a criterion of normative assessment for practical reasoning, or as specifying a constitutive aim of practical reasoning. Read the first way, perhaps it is a non-starter. Offering the best action as the criterion of normative assessment for practical reasoning is not as transparently lame as offering being a good friend as the criterion for friendship. Still, when we look into what is meant by "best action," the circle may seem to close. "Best action" may turn out to mean "action to which good reasoning would lead." But what we have just learned from our

discussion of friendship is that nothing whatever follows about the claim that in practical reasoning one aims at the best action. Let the circle close as it may, there will still be such a thing as reasoning well about what to do, and hence a best thing to do, just as there still is such a thing as being a good friend. To say the reasoner aims at this is to specify a genuine target.

Of course, this doesn't give us a criterion of normative assessment for practical reasoning. Velleman might hold that if there is to be such a thing as reasoning well about what to do, then there must be such a criterion. If this means that in each individual case, where a person reasons well or badly, there must be facts that show the reasoning to have gone well or badly, then I think we should agree. But in Velleman's hands it appears to mean something else, namely that it should be possible, in advance of particular cases, to fashion a master criterion that will separate wheat from chaff regardless of what turns up on the threshing room floor. This is what Velleman thinks he has, for the case of belief, in truth. Truth will serve as the master key that will fit every instance of reasoning about what is the case.

It is certainly part of the business of philosophy to look for the great master criterion of practical reasoning, but it is worth asking what it would mean if there were none to be found. Suppose the desired *a priori* test just isn't accessible to philosophy, no matter how well the philosophers go about their work. Suppose all we can do is fashion general principles that hold "for the most part," and then wait for the world to toss up what it will, before doing more philosophy about the particular case. What is the argument that, in that event, there is no such thing as reasoning well about what to do? Or are we just assuming that whatever philosophy can't cage doesn't exist?

Compare, once again, the case of friendship. What being a good friend amounts to is certainly a matter of philosophical interest. The philosopher will find marks, such as loyalty, well-wishing, curiosity about the other's thoughts and feelings, a desire to associate, and so on. But most of these things can be overdone or misdirected. Consider the prospects for finding any abstract formula that will tell us when the good friend must be loyal and when not, when the good friend asks and when she holds her tongue, when the good friend rejoices and when she is concerned, what she expects and what she doesn't, etc. The search for the master criterion is worthwhile for two reasons. First, there might be one. Second, just having searched can make us wiser judges of particular cases. But nowhere in the philosopher's handbook does it say we must believe everything has a philosophically accessible master key.

Limitations of philosophical method can't show there is no such thing as conducting oneself well or badly qua friend, and neither can they show there is no such thing as reasoning well or badly about what to do.

So I don't think Velleman's worries should be contagious. We need not be worried by the fact that agents sometimes act without aiming to act well. We can still regard acting well as a goal agents pursue insofar as their powers of deliberation function properly in them. And we need not be worried that the good fails to specify a substantive goal for practical reasoners to pursue. We can say practical reasoners, insofar as they reason at all, aim to do that very activity well. As long as there is such a thing as doing it well, they have something to aim at.

All in all, then, I think we've made a fairly promising start on a reply to Hume. We have absolved ourselves of any responsibility to explain how beliefs can work by themselves. The temptation to save the objectivity of ethics by rejecting Hume's doctrine about beliefs is simply misplaced. Objectivists like Nagel and McDowell feel driven to postulate beliefs that work alone, because they want to preserve the motivational magnetism of ethical beliefs. But if we work from within the theory of practical reasoning, we can use a goal of reasoning to explain the motivational magnetism. Hume's doctrine is not the enemy.

The real enemy is an impoverished sense of our options in the theory of practical reasoning. If we limit ourselves to the two familiar models of practical reasoning, there will be no goal of reason that can explain how objective ethical truths grip the will. One model has us aiming to discover ethical truths, and leaves us to cast about for an explanation of the magnetism of the resulting beliefs. The other has us aiming to fit our behavior to our desires, and gives us our internalism only if we are willing to build desires into the analysis of ethical beliefs. But there is a much richer model, on which the goal of practical reasoning is to make the steps that deserve to win. Like the goal of discerning the ethical truth, this goal can be harnessed without building facts about the agent's wants into the truth conditions of ethical beliefs. But unlike the goal of truth, this goal can press all the way home to action. It doesn't stall out in the formation of beliefs. Moreover, this model makes a place for both truth-oriented reasoning and the phenomenon of choosing an action for its connection with something wanted. The truth-oriented reasoning will be about which steps deserve to win, or where the weight of reasons lies. But making the steps will be choosing an action for its connection with something wanted. So rather than supplanting the old models with something entirely new, we treat the old models as parts of the correct story about the nature of practical thought.

Hume was right to wonder how reason could be practical. But that being the question, our only recourse is to say "Here's how," and get to work. That is what I have tried to do here.

Acknowledgements

This paper was written for a conference in memory of Warren Quinn, held at UCLA in October of 1998. Like most of my work, it feels incomplete without Warren's comments. For helpful criticisms and suggestions I wish to thank the audience on that occasion, the faculty and graduate students of the Yale Department of Philosophy, and Sergio Tenenbaum.

REFERENCES

Aristotle (1993). *The Nicomachean Ethics*. London: Penguin Classics.

Broadie, S. (1991). *Ethics with Aristotle*. Oxford: Oxford University Press.

Clark, P. (1997). Practical Steps and Reasons for Action. *Canadian Journal of Philosophy* **27** (1), 17-45.

Clark, P. (2000). What Goes without Saying in Metaethics. *Philosophy and Phenomenological Research* **60** (2), 357-379.

Clark, P. (2001a). The Action as Conclusion. *Canadian Journal of Philosophy* **31** (4), 481-506.

Clark, P. (2001b). Velleman's Autonomism. *Ethics* **111**, 580-593.

Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell.

Davidson, D. (1980). How is Weakness of the Will Possible? In: *Essays on Actions and Events*, pp. 21-42. Oxford: Clarendon Press.

Hume, D. (1888). *A Treatise of Human Nature*. Oxford: Clarendon Press.

Lawrence, G. (1995). The Rationality of Morality. In: R. Hursthouse, G. Lawrence and W. Quinn (eds.), *Virtues and Reasons: Philippa Foot and Moral Theory*, pp. 89-147. Oxford: Clarendon Press.

Little, M.O. (1997). Virtue as Knowledge: Objections from the Philosophy of Mind. *Nous* **31**, 59-79.

McDowell, J. (1978). Are Moral Requirements Hypothetical Imperatives? *The Aristotelian Society Supplementary Volume* **52**, 13-29.

McNaughton, D. (1988). *Moral Vision*. Oxford: Basil Blackwell.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton: Princeton University Press.

Platts, M. (1979). *Ways of Meaning*. Cambridge, MA: The MIT Press.

Platts, M. (1981). Moral Reality and the End of Desire. In: *Reference, Truth and Reality: Essays on the Philosophy of Language*, pp. 69-82. London: Routledge and Kegan Paul.

Stevenson, C.L. (1937). The Emotive Meaning of Ethical Terms. *Mind* **46**, 14-31.

Stocker, M. (1979). Desiring the Bad: An Essay in Moral Psychology. *Journal of Philosophy* **76**, 738-753.

Velleman, D.J. (1992a). What Happens When Someone Acts? *Mind* **101** (403), 461-481.
Velleman, D.J. (1992b). The Guise of the Good. *Nous* **26**, (1), 3-26.
Velleman, D.J. (1996). The Possibility of Practical Reason. *Ethics* **106** (4), 694-726.
Williams, B. (1970). Internal and External Reasons. In: *Moral Luck: Philosophical Papers 1973-1980*, pp. 101-114. Cambridge: Cambridge University Press.

Connie S. Rosati

MORTALITY, AGENCY, AND REGRET

We are creatures who are subject to regret. We survey past decisions about our educations and careers; we look back upon our friendships, family, and love relationships, sustained, developed, or fractured; we reassess opportunities seized or missed, talents and traits nurtured or neglected. When we reflect in these ways, we not infrequently judge that we have made mistakes: we judge that we did not choose what would have contributed to a good life for us. Part of what motivates efforts to arrive at a theoretical account of our good is surely our susceptibility to regret. We seek to determine how to live our lives and, thereby, how to avoid the regrettable, that of which our regret is a reflection.

Regret is commonly characterized as a particular cognitive and affective syndrome occasioned by the occurrence of an event or the coming about of a state of affairs that a person regards as undesirable.[1] The syndrome can involve various feelings, such as disgust, anger, or sadness, where these occur together with replays of events, choices, and actions, the thought that things could have been otherwise, and imaginings of how they might have been otherwise. Although the feelings involved in the syndrome may vary, the central cognitive component of regret remains fixed: our feelings are accompanied by the normative judgment that what has happened is in some respect undesirable or mistaken.[2]

This analysis of regret no doubt needs careful refinement, but the syndrome it depicts is familiar enough, and so in what follows, I shall

[1] I borrow the rough characterizations of regret and agent-regret contained in this paragraph and the next from Williams (1981) and Rorty (1980).

[2] See Rorty (1980, pp. 491-495), for discussion of the cognitive component of regret. As I shall suggest later, regret need not involve the judgment that what one has done or allowed to occur is contrary to one's good or mistaken all things considered. In this way, regret differs from remorse, which does seem to involve the judgment that one has made a mistake all things considered. See Baron (1988) and Zoch (1986).

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 231-259. Amsterdam/New York, NY: Rodopi, 2007.

simply assume that the analysis is roughly correct. I shall be interested solely in nonmoral regret or regret related to events or states of affairs regarded not as morally undesirable but as undesirable from the standpoint of an individual's good or well-being.[3] And I shall be interested primarily in agent-regret, or a person's regret about her contribution to an undesirable event or state of affairs through her own choices, actions, or character.[4]

Two central issues arise about regret. One issue focuses on the cognitive side of regret and concerns whether regret is ever well-grounded. The well-groundedness of regret depends on whether the normative assessments it involves are true or warranted. It depends on whether we rightly judge that we have acted contrary to our good or on whether we rightly conclude that we have brought about something undesirable. The second issue looks more to the affective side of regret and concerns whether regret is ever appropriate. The appropriateness of regret depends on norms about the rationality of emotions – about what it makes sense to feel. Such norms might dictate any number of relationships between well-groundedness and feelings of regret. They might tell us, for instance, that regret cannot be appropriate unless a person's normative assessment of her choice is warranted. Or they might

[3] When I talk about a person's good, my interest lies with the generic idea of what makes a life go well for the person who is living it. Various terms have been used to express this idea, most commonly, 'well-being', 'welfare', 'self-interest', and 'flourishing'. I have explained elsewhere my preference for the expression 'personal good'. See Rosati (2006). I suspect that these sundry terms have varying connotations. Though I tend to use them interchangeably, in this article I talk almost exclusively in terms of a person's good.

[4] From this point on, I will largely use the term 'regret' in place of 'agent-regret'. The context should make clear when my remarks apply specifically to agent-regret and when they apply to nonmoral regret more generally. As far as I have been able to discover, regret has received little attention in discussions of nonmoral good. Some theorists have at least implicitly appealed to our experience of regret and related emotions in defending their accounts of our good. For instance, Peter Railton asks us to consider the case of Beth, a happy and successful accountant, who gives up her career to pursue writing, for which (it turns out) she has no aptitude. He argues that his informed-desire analysis captures the normativity of a person's good by noting that the "all things considered" desires of the "sadder but wiser" Beth would weigh with actual Beth. See Railton (1986b, p. 13). John Rawls treats a person's good as given by her rational plan of life and argues that a rational person will not come to regret following an objectively rational plan, one chosen with accurate information and relevantly complete understanding of the consequences, even though that plan may not be "a good one judged absolutely." She may regret that her circumstances made a happier life impossible but not that she followed the best plan possible in her unhappy circumstances. A rational person may regret following a subjectively rational plan, but not because she regards herself as at fault in doing so. See Rawls (1971, pp. 416-424).

tell us that regret is inappropriate even when well-grounded; one might maintain of regret, as Spinoza observed of repentance, that it never makes sense to feel regret because all it does is make us "doubly wretched" (Spinoza 1951, Part IV, Prop. LIV).[5]

My discussion in this essay may have implications for an account of when regret is appropriate, but my interest will lie with the well-groundedness of regret. More precisely, it will lie with the very possibility of well-grounded regret.

1.

Now it might seem that we can easily explain how well-grounded regret is possible if we just consider for a moment the character of our regrets. When we reflect ruefully on a particular choice, we sometimes judge simply that the choice itself had unwelcome results or that it was mistaken relative to our other aims. But we often judge more deeply that we have made a mistake about our lives. In certain cases, we view our choices as permanently depressing the value of our lives. In extreme cases, we may say that we have "ruined" our lives. The negative judgments that accompany our feelings of regret in such situations reflect the same basic presuppositions that lie behind the judgments we commonly make when our choices please or satisfy us. In the latter cases, we often feel not merely satisfied but that we have made the right choice. We commonly describe our lovers or spouses as "made for us." We talk about "finding our niche." We claim to have ended up in the place we were "meant to be."[6] Our regrets, like our satisfactions, often seem to assume that we each have a determinate good against which our choices can be measured.

These considerations about the character of our regrets suggest a natural and seemingly plausible explanation of well-grounded regret: well-grounded regret is possible because we have a good which we can fail to discern. It is determined independently of our choices, beliefs, and evaluations, any of which may be mistaken. Our good is there to be

---

[5] Even if one rejects this extreme view, one might nevertheless think regret is largely counterproductive. For discussion and defense of the appropriateness of agent-regret in cases of moral conflict, see Williams (1973b). For an opposing (and more "Spinozistic") view to the effect that regret, in the moral case, is inappropriate because unreasonable, see Bittner (1992).

[6] In fact, the common tendency for people to talk about their lives in terms of what was "meant to be" is quite striking.

discovered – it provides the rails along which our choices should run.[7] When we feel regret, we judge that we have made a mistake about our good. Our specific regrets thus arise from perceived or actual failures to track our good, and well-grounded regret accurately registers just such a failure. It is worth noticing that this explanation implies that we could, in principle, eliminate regret from our lives. If a person could only discern her good and choose correctly, well-grounded regret would be avoidable; so would appropriate regret, at least insofar as well-grounded regret is ever appropriate.

This explanation of well-grounded regret captures well the intuition that a good life is a life without cause for regret, and I share that intuition, at least properly understood. But the explanation surely involves too simplistic a view of our lives and our good. Some of the judgments on which it rests, after all, clearly presuppose an implausible view of how the value of a life is determined, for they suggest that a life's value is irretrievably fixed by each discrete choice. The value of a life is affected by incorrect choice in much the way that the taste of a cake is affected by mishaps in baking – the taste is always off, for instance, because you forgot to add the vanilla, or the cake is ruined because you added salt instead of sugar. To be sure, the value of a life can be permanently depressed and lives can be ruined, but for reasons that will emerge later, this is not generally the case.[8] In any event, while many of our specific regrets may register perceived or actual departures from our good, not all of them do, at least not in the crude way that has been suggested.

Still, I don't doubt that an adequate account of the possibility of well-grounded regret must appeal in some suitably complex way to our good. I suspect that the underlying difficulty with the explanation we have just considered, then, is not that it links well-grounded regret to our good but

[7] The metaphor is Wittgenstein's, of course. See Wittgenstein (1953, Section 218), for his discussion of the picture of rules as rails. We might think of this as a folk picture of our good and our regret. It arguably has rough theoretical counterparts in perfectionist and objective list theories and, perhaps, in some informed desire theories, but my aim is not to saddle any particular welfare theorist with this simple view. Indeed, I suspect that many theorists could accept my explanatory story about regret, but if I am right, their theories of personal good have not adequately attended to the implications of that story for the nature of our good.

[8] I have argued elsewhere that certain things really are not dispensable or fungible if a person is to fare well. See Rosati (manuscript). When those things are missing, we might plausibly say that a life has been ruined. My point here is simply that it is not generally the case that making suboptimal choices either ruins or permanently depresses the value of a life. For a more detailed critique of the idea that the value of a life is additive, see Velleman (1991).

that it is missing a critical first step. We are creatures who are subject to regret – both well- and ill-founded – and since our good must fit the sort of creature we are, the structure of our good itself presumably reflects this fact about us. For this reason, I am inclined to think that even if we must appeal to our good in order to understand the possibility of well-grounded regret, we must first understand why we are given to feelings of regret if we are to understand fully the nature of our good. Once we appreciate why we are subject to regret – why we experience it at all – we can gain insight into the nature of our good and, thereby, into the possibility of well-grounded regret.

In the discussion that follows, I hope to lend support to this idea and, ultimately, to reorient our thinking about a person's good. To that end, I offer an explanatory story about why we feel regret, exploring what it helps to reveal about the nature of our good. According to my account, our experience of regret is most deeply explicable in terms of our unique features and circumstances – in particular, our character as persons or autonomous agents. Those aspects of our character and circumstances that explain why we are subject to regret also explain why we must develop a conception of our good and why our good must assume the shape that it does. More deeply, they explain the peculiar respect in which even well-grounded regret is, in the end, unavoidable. Our good, I will suggest, is best understood as a solution to a certain, practical problem we inescapably confront in living our lives. The moral of my story will be that, given the nature of our good, the possibility of well-grounded regret rests, in fundamental ways, with us.

2.

Before we turn to my account, it is worth examining our regrets more closely. A closer inspection of them will illuminate what we are subject to in being subject to regret and, thus, what needs to be explained by an account of why we experience regret. Not only do we feel agent-regret and judge our choices undesirable or mistaken, but our feelings of regret and our assessments of our choices are liable to change. Our evaluations of our decisions and aims undergo curious shifts, as do our feelings of regret, and these will require explanation.

First, our regrets and our judgments of our choices as mistaken are susceptible to a phenomenon akin to the one Bernard Williams and Thomas Nagel have called "moral luck" with respect to rational

justification and judgments of moral responsibility.[9] Aristotle considers a phenomenon like the one I have in mind in the *Nicomachean Ethics* when he discusses how changes of fortune affect whether we call a man happy. Happiness, he tells us, requires "not only complete virtue but also a complete life, since many changes occur in life, and all manner of chances, and the most prosperous may fall into great misfortunes in old age, as is told of Priam in the Trojan Cycle; and one who has experienced such chances and has ended wretchedly no one calls happy" (Aristotle 1926, 1100a). Aristotle was understandably bothered by the implications of shifting our judgments of a man's happiness in accordance with changes in his fortune, for it makes "the happy man out to be 'a chameleon, and insecurely based'" (Aristotle 1926, 1100b). Our ordinary conception of happiness, Aristotle tells us, is of something more permanent and stable. His own characterization, he contends, provides the needed stability; the happy man, since he continually engages in activities in accordance with virtue, will nobly withstand changes in his fortune. Aristotle acknowledges, though, that while the minor good and ill effects of fortune cannot undermine a man's happiness, his bliss can be increased by great successes or diminished by severe misfortunes.

> Yet even in [adversity] nobility shines through, when a man bears with resignation many great misfortunes, not through insensibility to pain but through nobility and greatness of soul. If activities are, as we said, what determines the character of life, no blessed man can become miserable [. . .]. (Aristotle 1926, 1100b)[10]

Aristotle was surely right that some persons – including perhaps those virtuous in just his sense – well resist adversity. Even such persons, however, typically regard their choices as contributing well or ill to their good. And whether they regard their choices as mistaken or experience regret about them is often due to circumstances beyond their control. If John spends his last fifty dollars on lottery tickets and wins, he will most

[9] Williams and Nagel do not take themselves to be talking about precisely the same phenomenon. See Nagel (1979, fn. 3) and Williams (1981, pp. 36-39). But each discusses how our regrets and evaluative judgments are affected by how our choices and actions turn out.

[10] Aristotle's talk of the happy man's nobility shining through adversity is reminiscent of Kant's remarks about the good will, sparkling like a jewel in spite of any inability to accomplish its purposes. See Kant (1964, p. 62 [Prussian Academy edition, p. 394]). Apparently, Kant thus sought to insulate our moral judgments of persons and their actions from the effects of fortune. (Nagel 1979, p. 24, makes this point.) Aristotle's discussion suggests a similar attempt to insulate the happy man and our judgments of a man's happiness from the effects of fortune, though he does not go so far as Kant, admitting that good and bad fortune can affect a happy man's bliss.

likely judge that his choice was not mistaken and be without regret for his action. If he loses, though, he will surely regret having squandered his resources. He may criticize his own action as imprudent – or more strongly, as foolhardy – even if he wins. But since it is crucially through his own agency that he wins, he is not likely to be sorry that he acted as he did.[11] If Ron decides to devote all his time and money to ballet lessons to the detriment of his other interests, whether he judges his choice mistaken will depend in part on how things turn out. Should he succeed in his goals, he may regard the time and money as well spent; failure, however, may occasion regret and his retrospective evaluation of his choice as mistaken.

How we judge our choices, then, is affected in part by factors beyond our control. Luck affects our assessments of our choices because, as Williams and Nagel suggest, what we have done and, hence, the significance of our doings, depends in part on how things turn out.

Yet the changes in our regrets do not neatly follow changes in our fortune; they are not merely a reflection of how our choices fare. On the contrary, people often do not regret unsuccessful choices, while strongly regretting ones that successfully realized the very aims they had set out to achieve. A person can regret even a successful career in sales, for example, having chosen it over the priesthood. While some changes in our regrets mirror our shifting fortunes, as a general matter, our regrets may shift depending just on when and how we look at our choices. We commonly regret and regard as mistaken at one time choices that we do not regret or regard as mistaken at another time. Sometimes we come to regard an earlier choice as mistaken and acquire regret; in other cases, we revise our negative assessment of an earlier choice and lose our regret.

Changes of the first sort are common enough. People often end relationships, only to feel pangs of regret at having given them up years down the road. Sometimes people later regret having made particular career choices, especially as they grow older and realize that they have closed off options that once attracted them. And sometimes they later regret not having pursued various projects – they wish that they had been more politically or socially active, or that they had learned more about the arts or history – even though they did not regard failure to pursue such projects as mistaken at the time.

---

[11] This is not to deny that John might continue to have other regrets surrounding his choice, such as regrets about his character. One might think that even though John's choice turned out well he should feel regret. But that is a point about the appropriateness of regret, not about whether he will in fact feel it.

Changes also commonly occur in the other direction. People often regard relationships they have had as mistakes. Immediately after a painful breakup of a problematic relationship, a person may regret her involvement in that relationship. Over time, however, she may lose her regret and even her judgment that the involvement was mistaken, although she retains the belief that the relationship was problematic. Similarly a person may regret not having taken advantage of a particular career opportunity, only later to regard it as an opportunity well enough missed.

3.

How are we to explain the susceptibility of our judgments of mistaken choice and our feelings of regret to shifts, depending upon how things turn out and upon when and how we reflect on our choices? Certain natural explanations suggest themselves, and these explanations will often have force with regard to our regrets in specific cases. But they do not fully explain our changing regrets.

According to a first explanation, perhaps best illustrated by cases of "nonmoral luck," whether we feel regret about aims and actions depends upon whether we believe that we would have undertaken them, had we contemplated them rationally and in full non-evaluative knowledge of how they would fare. We typically lack complete information about how an aim or action would turn out when we deliberate about whether to undertake it. Moreover, we are less than fully rational choosers; even when we have adequate information, we tend to make mistakes in instrumental reasoning and are liable to various cognitive shortcomings. We want to make choices that are well-informed and rational. Our assessments shift when we come to believe that we have failed to do so.[12]

Now sometimes, perhaps often, our regret arises because we operate with limited information and rationality. But our epistemic and cognitive limitations do not fully explain the shifts in our regret. For sometimes we may regret choices even though we believe that, had we known more and deliberated rationally, we would not have chosen differently. And conversely, sometimes we do not regret what we think we would, knowing more and thinking rationally, have chosen to avoid. Sometimes

[12] This first explanation might be offered by informed-desire theorists and is suggested by discussion in Rawls (1971, p. 422), Railton (1986b, p. 13), and Brandt (1979, pp. 154-160). But obviously the explanation is a common-sense one, requiring no particular theoretical commitments.

we are even glad we did not know how things would turn out, because then we might not have chosen as we did. Ron may not regret his decision to pursue dance, for instance, even though he believes that, had he considered fully and rationally how his aim would fare, he would have chosen to spend his time differently.

Our regrets can diverge in these ways from what we think our fully informed choices would have been because we need not think our fully informed choices settle how to assess our actual decisions. How we would react and choose, after all, does not depend solely on information and rationality. It depends as well on our current motivational systems, and we do not invariably trust that our own motivational systems pick out what is worth desiring, even when infused with full non-evaluative information and rationality.[13] Thus, while limited information and rationality explain some changes in our regrets, they will not explain them all.

A second explanation accounts for our shifting regrets as involving two separable judgments, with the second judgment coming to replace or modify an earlier one. For instance, a person might at one time regard a choice as objectively mistaken and feel regret; later, however, she may come to see it as subjectively unmistaken and so lose her regret. She may still believe that she ought not to have made the choice, but she gives herself credit for having chosen as well as she could have in the circumstances, and in this respect she does not regret doing what she did. In such cases, what a person really does is retain her negative assessment but release herself from self-reproach.[14] The person who regrets a problematic relationship, for instance, may not blame herself, if she had reasons at the time to believe the problems could be resolved.

This explanation correctly observes that we may simultaneously view actions from different standpoints, standpoints that lead us to make

---

[13] We need not be prepared to allow what is worth desiring to be settled causally through the workings of our own motivational systems, even when corrected for mistakes in reasoning and information. Gibbard makes a related point with respect to full information analyses of 'rational' (1990, pp. 21-22). My remarks here are obviously not intended to rebut full-information accounts of a person's good but merely to show that our regrets are not fully explained by appealing to departure of our actual choices from what we think our fully informed desires might be. For critical discussion of full-information theories, see Velleman (1988), Sobel (1994), Loeb (1995), and Rosati (1995a, 1995b).

[14] As Rawls has suggested, for instance, we need to distinguish between a person's objectively rational plan of life, which reveals her true good, and her subjectively rational plan of life, which represents the best she can do following principles of rationality and working with limited knowledge. If she follows her subjectively rational plan, Rawls says, a person need never feel self-reproach (1971, p. 422).

distinct judgments about them. It also rightly reminds us that when we believe that we chose carefully, we may be less likely to criticize ourselves or to regret having chosen as we did. Yet a person apparently can regard a choice as both mistaken and not mistaken or as both desirable and undesirable, and not because she regards it as objectively, but not subjectively, mistaken. The person who has chosen to forego having a family for the sake of pursuing her scientific research need not think the choice of either of those options would have been an objective mistake.[15]

According to another version of this "two judgments" explanation, our regrets change because we shift from an intrinsic to an extrinsic view of our choices.[16] When we revise our assessments and later judge as unmistaken a choice we formerly regarded as mistaken, it is because we see that we have learned something from the experience or that it has changed us in desirable ways. Thus, the person who later ceases to regret a problematic relationship has come to view it instrumentally; viewing it intrinsically, however, she must still think it mistaken. Insofar as we judge a choice mistaken or somehow undesirable, we judge it as not to be preferred for its own sake. But even in the midst of a difficult experience, we can recognize its contribution to our development, our learning, and our acquisition of other desirable things.

This explanation correctly reminds us (not that we needed reminding) that we do not in general like pain for its own sake, and insofar as experiences are painful, we tend to regard them as mistaken. Nevertheless, as the explanation shows, we appreciate the instrumental value of even painful experiences; we understand that certain kinds of growth and insight have their price.

Yet this explanation, too, is not fully adequate. Because our choices bring about complex results, it is doubtful that they can in all cases be neatly or meaningfully or finally categorized as intrinsically or extrinsically mistaken or unmistaken. In any case, we do not assess our choices only intrinsically or instrumentally; we may also assess them in

---

[15] Joseph Raz has expressed the point as follows: "On the not very frequent occasions when one may choose between different goals which are not as yet one's own one may well not have a decisive reason to choose either way." Thus, he says, I may not have a reason to prefer going to medical school rather than law school. When I choose the former, I will have reasons for doing so, "and the fact that I believe that I also have reasons for choosing a legal career which are no less worthy and important does not undermine the fact that when I choose medicine I choose it for the stated reasons" (1986, p. 304).

[16] Or vice versa. I here examine only the more intuitive side of the equation, but the same difficulties apply to the other side.

terms of how they contribute to the value of a whole, for instance, to a life story or segment of that story, without seeing them as merely instrumental.[17] Indeed, whether we assess a choice intrinsically or instrumentally depends upon the larger context in which we view it. For instance, Ron may regard his choice to pursue dance as intrinsically mistaken but instrumentally valuable if he views it as part of a life in which the effort he invested paid off later in a different career. Or he may regard the choice as intrinsically unmistaken if he views it as a part of a life story in which he was free to spend his youth in artistic endeavor.

Our shifting regrets thus need not be due to a lack of information and rationality or to separate but compatible judgments of the sorts we have been considering. Of course, the foregoing explanations by no means exhaust the possibilities, and as I have already allowed, explanations like these may often adequately explain our specific regrets. But we will gain a more complete explanation of the vicissitudes of our regret once we understand why we experience regret at all.

4.

We can begin to uncover the underlying source of regret by examining a type of case now commonplace from discussions of personal identity, namely, split-brain or fission cases. In certain split-brain cases, we are asked to imagine that the hemispheres of our brains are symmetrical.[18] Our brains will be divided and each hemisphere placed in a new body. We are then asked to consider whether we believe we would personally survive the imagined procedure. These science fiction cases are of interest, not only for what they display about our notion of personal identity, but for what they reveal about our ordinary concerns in living a life. We can thus make use of fission cases, as I will here, as a device for bringing into relief these ordinary concerns. In so doing, I suggest, we will also bring into relief those circumstances that give rise to our regret.

Let's focus on a fission case that sidesteps not only worries about differing capacities of the two hemispheres of a divided brain, but worries that arise when split brains are placed in bodies differing from those in which the intact brains were originally housed. Imagine an injection that causes a person to divide like an amoeba, each half growing the missing parts, including the missing side of the brain. Both of the

[17] For two quite different discussions of the limitations of the ends-means distinction, see Kolnai (1962) and Korsgaard (1983).

[18] See Parfit (1973; 1975; and 1984, Part 3).

resulting persons are psychologically continuous with the person who underwent fission; each has strong relations of psychological connectedness – of memory, intention, and expressions of traits of character – with that person.[19] But each will go on to lead a different life, to have different experiences, to acquire different memories, and to change physically and emotionally as a result of her choices.

Our worries about whether, were we divided, one, both, or neither of the resulting persons would be us do not arise simply because we are torn between conflicting demands of our ordinary criteria of personal identity, between psychological and physical continuity on the one hand and numerical identity on the other. Rather, they arise because we are well aware of the sundry changes that will occur after the splitting has taken place. As a consequence, we tend, fittingly, to be of two minds about such cases. On the one hand, we have the uneasy feeling that the resulting persons cannot both be us, and not simply because one does not equal two. They cannot both be us because we feel and live our lives from a particular point of view, and here are two points of view which, however alike at the beginning, will become increasingly different. One or the other or both might go on to develop in ways that we would not like, with which we cannot now identify. On the other hand, we can relish the prospect of being able to live more lives than one, for then we would not have to choose between our conflicting desires or between the different lives that attract us. We could, for instance, both live our lives as philosophers and live our lives doing that other thing we always thought we might do when we felt like quitting philosophy.

We are attracted to the prospect of leading more than one life because of features of our ordinary circumstances. Just as we operate under what John Rawls has called the "circumstances of justice," which make possible and necessitate that we embrace a vision of a social order and a conception of justice, so we operate under "circumstances of the good," which condition our feelings of regret and our judgments of mistaken choice, and which necessitate that we embrace a vision of a life and a conception of our good.[20]

The circumstances of justice, according to Rawls, include both objective and subjective circumstances. The objective circumstances of justice include moderate scarcity of resources and the coexistence of and attendant conflicts among people with roughly comparable physical and

[19] I here follow Parfit's characterization of psychological continuity, though I have presented it in a simplified form. See Parfit (1975, pp. 214-215; 1984, pp. 205-209).

[20] Rawls (1971, pp. 126-130). The circumstances of justice, Rawls says, are the "normal conditions under which human cooperation is both possible and necessary" (1971, p. 126).

mental capacities. The subjective circumstances of justice include that persons each have their own plans of life involving different aims and ends that compete for available resources, that persons have no particular interest in each other's interests, and that they operate with various cognitive and informational shortcomings. These conditions make necessary principles that resolve conflicts among persons so that they can lead reasonable lives and fulfill their plans at all.

The circumstances of the good can similarly be divided into objective and subjective circumstances. The objective ones include that we operate under conditions of material and temporal scarcity. We are forced to work with limited material resources – chief among them, that we are limited to just one life. Moreover, we are limited temporally – we not only have just one life to live but a short one at that.

The subjective circumstances include that we operate under conditions of limited information and rationality and that we are subject to pain. In addition, we each have conflicting desires, all of which we would like to see satisfied, for part of having desires is having some motivation to see them satisfied. This motivation is not experienced, however, as a bare push or pull. Desire is itself a partly evaluative state, so that when we desire something, we find ourselves drawn to what we see as at least in some respect desirable.[21] The conflict of desires that we experience is thus a conflict among attractions to things each of which strikes us as the thing to pursue. It is a conflict, moreover, among desires that belong not only to us, but to the different persons whom we could become and who compete for our attention and our material and temporal resources; our conflicting desires include ones the choice of which to satisfy will commit us to becoming different sorts of people with differing aims and motivations, each of which may attract us.

Faced with these sundry desires and limited resources, restricted as we are to a single life of limited duration, we are pressed psychologically to choose among our desires in order to satisfy any of them at all; indecision would leave us losers all around. We live our lives, moreover, as persons or autonomous agents. We are creatures with a self-reflective awareness of our choosing, the ability to interpret our choices and experiences, and a concern for what we are like that manifests itself in the formation of higher-order desires. As consequence, we are aware of

[21] This is not to say that a desire is a value judgment in the sense that the constitutive aim of desire is getting right what is worth having or bringing about. And so the constitutive aim of desire is also not getting right one's own good. It seems one can desire the bad as such, including what is bad for oneself. See Stocker (1979) and Velleman (1992). Compare Stampe (1987).

our choosing and aware that in choosing among our desires we give up something else that we wanted and might have had instead.[22] We realize, too, that certain choices will irreversibly limit our future options. The circumstances of the good give us motives to resolve conflicts within ourselves, thereby making it possible for us to lead a life and to have a good at all.[23] They also give rise to regret.

Let's examine more closely certain critical aspects of the circumstances of the good. Consider first our mortality.[24] We live our lives under conditions of temporal scarcity, and this temporal scarcity is of a complex sort. Not only do we have a limited lifespan, so that we cannot do or pursue all that we would like, but we also age, and pursuits that were possible at some points within the progress of a life are no longer possible at later points. Think about the impact of a woman's declining fertility as she ages – the winding down of her "biological clock," or the limitations on opportunities that may come with entering upon a particular career later in life. Thus, not only does limited time affect us, so does timing. Our choosing is conditioned, not only by the mortality that marks the end of a life but by a kind of mortality that arises repeatedly within a life.

At any period within our lives, we typically have many and conflicting desires. We might describe our position in relation to our desires as Mae West described her lamentable position in relation to men: so many desires, so little time. Just as when we have limited money, we wonder which of those things we desire to spend money obtaining, so when we have limited time, we wonder which of our desired ends to spend time pursuing. In either case, we are driven to decide how to expend our resources, partly by our desires themselves. Our desires

---

[22] Think of the old Lovin' Spoonful song, "Did You Ever Have to Make up Your Mind?".

[23] See Korsgaard (1989). Korsgaard's discussion of personal identity has informed my thinking about the relationship between regret and our good. Korsgaard argues that the unity of agency is pragmatically rather than metaphysically based. It is rooted in the necessity of eliminating conflicts between our motives, together with the fact that as agents, we regard ourselves as having reasons for choosing among our conflicting motives. The unity of agency thus arises from the necessity of eliminating conflicts between our motives, together with a unity already implicit in the standpoint we occupy as agents who deliberate and choose. Together these two elements impose upon us the necessity of identifying with a way of choosing among conflicting desires. See Korsgaard (1989, pp. 110-115).

[24] Discussions with David Reed-Maxfield and Michael Winstrom have led me to appreciate the importance of our mortality for how we live our lives, though I probably do not interpret its importance in the same way that they do. For reflections on our mortality and an argument to the effect that it is a good thing that we are not immortal, see Williams (1973a).

themselves give us motives for choosing because having a desire partly involves having a motive to satisfy that very desire. Since we cannot satisfy all of them, but have a motive to satisfy each of them, we are pressed to resolve the conflict, thereby satisfying at least some of them. Our conflicting desires together with our limited temporal resources force us to choose.[25]

We must exercise care in how we choose, however, for we operate with limited material resources, having only one life to live. Our possibilities are restricted not only because we have finite time and so cannot do all that we would like but because certain choices permanently foreclose other possibilities. If we make a choice that limits our later possibilities, ones that we may come to see as desirable either because of changes in our desires or because of changes in ourselves and who we have come to be, we cannot "wipe the slate clean" and begin a new life.

Consider next the sort of creature we are. We are persons or autonomous agents. Just what our autonomy consists in is a difficult question. But whatever the proper analysis of autonomy might be, autonomy will involve successful exercise of those capacities that render us self-governing, that help to free us from the immediate grip of our desires so that we are not simply moved by whichever first-order desire is presently strongest. The relevant capacities may include the capacity for self-understanding, that is, the ability to describe and explain what we are doing, as well as the capacities to engage in self-reflection, to exercise imagination, to reason and be moved by reasons, and to form and act on higher-order desires.[26] The successful exercise of these

---

[25] The difficulty is not really mitigated if we restrict our attention to desires for things that are themselves good in the sense that they can help to make a person's life go well, for there are far too many such things.

[26] The idea that certain motives and capacities are either constitutive of or at least essential to agency has been suggested by a number of writers. Velleman has argued (1989), that intrinsic desires for self-understanding and self-awareness, or more recently, an inclination toward autonomy (1996), are constitutive of agency. Velleman evidently believes that these characterizations come to roughly the same thing. See Velleman (1997, p. 41, n. 20). The motives constitutive of agency enable autonomous action, Velleman argues, by giving an impetus to our reflective choices as against our currently strongest lower-order motives. Michael Smith has argued that a disposition toward coherence is constitutive of rational agency. See Smith (1997). According to Smith's dispositional theory of value, facts about what we have reason to do are facts about what we would desire that we do insofar as we had the maximally coherent and unified set of desires that all persons would converge on under the influence of increasing, relevant information. Because rational persons have a disposition toward coherence, they tend to acquire desires that match their "evaluative beliefs," their beliefs about the desires they would have if they were fully rational and, moreover, to act on those desires as against whatever desire

capacities requires that we have some motivation to exercise them. I assume that the autonomy of autonomous agents depends on their having the desire to persist as the sort of creature they are, and so they will naturally develop concomitant, intrinsic desires to exercise those capacities the successful exercise of which renders them autonomous.[27] As autonomous agents, our capacities enable us to occupy a perspective from which we can experience a conflict of desires as such and face the question of what to choose. And because we are autonomous agents, we do not simply choose among conflicting desires: we choose self-reflectively, with a concern for the sort of person we will become and with an abiding interest in maintaining our autonomy.

Our experience of regret depends crucially on the very capacities that render us autonomous. Our ability to reason, to imagine, and to reflect on ourselves and our choices enables us to envision what our respective options might be like and what sort of people we might become as a result of choosing them. Because we reflect, we are aware when we make a choice of the options we did not choose. We are thus aware that in choosing, we close off other lives we might have had and other persons we might have become. Self-reflective choosing gives rise to feelings of loss. What's more, we not only reflect on our choices, we also interpret and reinterpret our choices and ourselves in our efforts at self-understanding. We can feel the loss of those lives we did not choose, not only at the time of choosing, but later in reflecting back on our lives, in reinterpreting our choices and our selves as the self we have become

---

may immediately be most pressing. Brandt has argued that humans happen to have standing desires for their own long-term happiness and for desires that are consonant with reality, and these standing desires enable them to act (against a present desire) in favor of their longer-term interests. See Brandt (1979, pp. 156-157 and 85). I have defended the idea that Brandt's work contains a picture of agents as equipped with certain standing motivations in Rosati (2000). Finally, Rawls has argued that the possession of certain moral powers (the capacity for an effective sense of justice and the capacity to construct, revise, and rationally pursue a conception of the good) and corresponding highest-order interests in exercising them is constitutive of persons on a Kantian ideal and renders persons autonomous in the original position. See Rawls (1980, p. 525). Rational agents in the original position are autonomous, he suggests, partly because they are motivated solely by these highest-order interests and are not required to follow or apply antecedent principles of justice (Rawls 1980, p. 528). The importance of capacities for self-reflection and the formation of higher order desires has been discussed by numerous writers. See Dworkin (1970), Frankfurt (1971), Neely (1974), and Watson (1975). See also Taylor (1985a), Nagel (1986, Ch. VII). For a critique of endorsement views of autonomy, see Buss (1994).

[27] We find expression of something akin to the idea that autonomous agents want to persist as such in John Stuart Mill's famous observation that a discontented Socrates wouldn't consent to become a happy fool.

through our choices. We lose not only the lives we did not choose as we once interpreted them, but the lives we did not choose as we now interpret them.

Consideration of the circumstances of the good enables us to explain why we are subject to regret: we are subject to regret because our unique nature and circumstances force us to choose reflectively among our desires and to experience the attendant loss that choosing involves. When I say that we experience regret because as autonomous beings we must choose reflectively among our desires and experience loss, I do not mean to suggest that what we always feel when we experience regret is sorrow about having had to choose. On the contrary, as I have noted, our regrets often involve something more than the judgment that our choice was in some way undesirable; they are, indeed, often inseparable from the feeling that we have chosen wrongly or contrary to our good, and this requires explanation. Nevertheless, I believe that we do experience a distinct "choice-regret" about having to choose and to lose an option we regarded as desirable, and such regret involves no sense that we have chosen mistakenly all things considered. In such cases, our regret expresses a kind of grief, a mourning of our lost lives and selves.[28]

Cases of nonmoral conflict, in which we cannot as a matter of fact realize both of the conflicting alternatives, thus resemble cases of moral conflict. In each kind of conflict, a person may choose and feel regret without thinking her choice mistaken all things considered. That is because both kinds of conflict are more akin, as Bernard Williams has argued in the moral case, to clashes of desire than to clashes of belief. As Williams observes, when a person who has conflicting beliefs becomes convinced of one of them, she gives up the other belief as false. But when a person who has conflicting desires chooses to satisfy one of them, the unsatisfied desire need not go away. Cases of moral conflict, he argues, work in just this way: the person who faces a dilemma may choose between conflicting moral demands and act in the belief that she has chosen for the best, without losing her belief that she has, in the process, done something that she ought not to have done. This belief may manifest itself, Williams suggests, as regret.[29] Similarly, when we must choose between possible selves and lives we regard as desirable, we are unlikely upon choosing, to lose our sense that we have given up

---

[28] I want to thank Paul Davies, who originally suggested to me that grief might best describe what I was after here. Our regret, in which we sense our good as closed, and our sense of our good as open, which is arguably manifested in Moore's "open question" argument, appear to be twin expressions of our agency.

[29] See Williams (1973b). Also see De Sousa (1974).

something valuable and appealing. Some of our regrets may thus persist quite apart from any sense that we have chosen wrongly or contrary to our good.[30]

Fission cases hold a certain fascination for us, I think, because they present a special case of our ordinary decisionmaking, one that not only reveals, but apparently circumvents, those conditions that occasion our regret. But whether one finds the prospect of fission attractive or repugnant, there is nothing unusual about the problem I use it to illustrate. Whether or not you would ever want to undergo fission, you have likely wanted to pursue more than one available option and wished that you could do both. In the ordinary case, as in the fission case, a person confronts the possibility of two (or more) different trajectories her life might follow, each with differing implications for who she will be and what her life will be like. Think about the person who has not yet decided between graduate school in philosophy and in business. What she knows is that she must choose between widely differing paths her life might follow, each of which will result in her acquiring different experiences and memories and developing different traits, interests, and values. But the difference between the ordinary case and the fission case is palpable: the person who elects to undergo fission makes a choice that apparently enables "her" to have both lives and be both people; in contrast, an actual person makes a choice that enables her to have only one life, to be only one person. The "fissioned" person's choice apparently secures her access to another self and life, whereas the actual person's choice closes off to her another self and life. Those of us who find the prospect of fission intriguing do so because it apparently presents us with the prospect of having our cake and eating it too, although closer examination of fission cases reminds us of why that is not possible. We apparently could have our cake and eat it too because, having more than one life to live, we would not be forced to choose between different lives: a person could live one life as a philosopher and the other as a businesswoman. Yet we cannot have our cake and eat it too, since there would, ultimately, be no us to have it both ways. There

[30] Thus, not all of our regrets involve the judgment that we have acted contrary to our good even though they involve the judgment that what has happened is in some respect undesirable. This conclusion is further supported by the obvious consideration that even sceptics have their regrets. I do not mean to deny that some individuals may fail to feel any regret unless they believe they have chosen mistakenly all things considered. But if nonmoral conflict is indeed like cases of conflicting desires, then few people will be wholly without choice-regret. Even those who believe that persons have an independently fixed good can allow that when an individual must choose between lives that tie for being her good she may experience choice-regret.

would, after all, be two persons, with differing points of view. And they would be two people confronted with the circumstances of the good and, consequently, subject to regret.

5.

Suppose that this account of why we experience regret is correct. What insight does it yield into our good and, thereby, into the possibility of well-grounded regret? I believe that the account helps to expose the essential nature of a person's good, and what it reveals should, accordingly, guide our efforts to construct theories of welfare.

Consider our position as creatures who must act and live out our lives in the circumstances of the good. As autonomous agents, beings with the capacity to reflect on our selves and our circumstances, we confront a practical problem. We have conflicting desires, each of which gives us a motive for its satisfaction, but given our limited temporal and material resources, we cannot satisfy them all. Our desires themselves, together with the circumstances in which we operate, press us to resolve conflicts among them: without coordination of our desires, few if any of them will be satisfied.[31] Moreover, without coordination of our desires, our ability to function autonomously is impaired and thus our most fundamental desires – the desires that help to constitute us as autonomous agents – go unmet. Our practical task is to form a coherent, stable, and attractive ordering of our possible aims. But that is just to develop a conception of our good.[32] We are impelled to devise a conception of our good, then, as a coordination device in the face of the circumstances of the good. Were we unable to order our aims in this way, it would not be possible for us to lead lives that are recognizably good at all. The circumstances of the good thus explain why we are subject to regret, and at the same time, why we come to develop a conception of our good.

Notice that a person's conception of her good plays a role for the individual akin to the role played by principles of justice for a group of individuals. We can see this by looking at a basic motivation for

[31] For extended discussion, in a different context, of practical reason and our need for intrapersonal coordination of our aims and activities, see Bratman (1987).

[32] I assume, of course, that we are not talking about aims or desires one has only insofar as one is concerned about the requirements of morality. I here roughly follow Rawls in treating a conception of the good as an ordered scheme of final ends, together with a story about what makes those ends appropriate or worthwhile, though I mean to refer to a person's idea as to a nonmorally good life for herself, whereas Rawls' idea has seeming moral elements which I want to leave to one side. See Rawls (1980, p. 544).

"constructivism" about justice.[33] Suppose that there were no "independent moral order," no moral facts or facts about justice fixed independently of what we might believe those facts to be. The circumstances that create the necessity for principles of justice would not disappear. Principles of justice respond to an obvious and pressing need, one that persists in the absence of independent moral facts: the need to coordinate the activities of and adjudicate conflicts among persons with competing commitments and desires who coexist under conditions of scarcity.[34] As constructivism about justice suggests, if there weren't principles of justice fixed by a "prior and independent order of objects and relations," we would have to invent them (Rawls 1980, p. 519).[35]

The foregoing explanatory story about regret suggests a parallel motivation for persons to adopt a conception of their good. Just as the circumstances of justice make it necessary for us to develop a conception of justice, whether or not there are independent moral facts, so the

---

[33] Of course, Rawls (1980) offers a far more rich and detailed rationale than the one I here sketch. As Rawls describes it, Kantian constructivism addresses a problem arising from the continued absence of an agreement "on the way basic social institutions should be arranged if they are to conform to the freedom and equality of citizens as moral persons" (p. 517). It holds that "apart from the procedure for constructing principles of justice, there are no moral facts" (p. 519). As Rawls explains it, the problem of justifying a conception of justice is practical rather than epistemological. The concern is to arrive at a "public and workable agreement on matters of social justice which suffices for fair and effective social cooperation" (p. 560). Compare Rawls (1980) with Rawls (1993, Lecture III, especially pp. 93-101), wherein the focus is on specifically political constructivism as contrasted with Kantian moral constructivism. Rawls describes political constructivism as neither denying nor asserting that there is an independent moral order (Rawls 1993, p. 95).

[34] One might argue that the existence of a rational demand for principles of justice and morality legitimates the move from merely accepting these principles to recognizing them as correct. I won't explore here whether there are any good grounds for accepting this move. See Grice (1991). Grice contends that the existence of a rational demand for absolute value does not merely justify the acceptance of certain attributions of value but grounds their truth. He recognizes the worry that transcendental arguments (such as the argument that he offers for absolute value) may show at most that rationality requires acceptance of a thesis. But he suggests that the demand for acceptance of a thesis may be sufficient to secure its truth when "we must accept such and such a thesis or else face an intolerable breakdown of rationality" (p. 106). I do not think, however, that Grice has shown that abandoning belief in absolute value would result in such a breakdown, and if certain noncognitivists are correct, it needn't.

[35] According to the constructivist, that is essentially what we do. David Brink has argued that constructivism is most plausibly thought of not as an alternative to but as a form of realism. See Brink (1987). We needn't broach the realism/anti-realism debate to appreciate the attractions of constructivism. We can view the latter not so much as providing a view in opposition to realism but as a view about the shape of facts about justice.

circumstances of the good make it necessary for us to develop a conception of our good whether or not we have a determinate good that is out there to be discovered. A person's conception of her good responds to an obvious and pressing need, one that persists in the absence of independent facts about her good: the need to resolve conflicts among her desires and her possible selves, as they arise under conditions of material and temporal scarcity. Only if this need is met can she be unified as an agent; only then will she be able to *lead* a life rather than be led by her desires; only if she can actually lead her life does she have a chance at a satisfying life. She must, accordingly, put herself in a position to do so.[36]

Attention to the role played by a person's conception of her good, I suggest, should inform our understanding of the essential character of a person's good: a person's good can be understood as that which addresses well the practical problem which she must meet. An adequate theory of our good must explain then not merely how we come to coordinate our aims and develop a conception of our good, but how, in virtue of our circumstances, it makes sense for us to do so.[37] Clearly we could not coordinate our desires in just any old way and successfully resolve the problem we confront; not just any efforts at ordering will produce a coherent, stable, and attractive set of aims. This is just to acknowledge the normative gap that may exist between a person's actual conception of her good and our concept of a person's good.[38] To close that gap, I want to suggest, is to order our desires in a way that fits our nature and circumstances.

Any theoretical account of our good that attempts to explain how this closing is achieved must attend to a number of important considerations. To begin with what is most obvious, not just any good that we may happen to desire, or might come to desire, can be good *for us*. We may lack the capacity to undertake some activities and pursuits with any

---

[36] The way this works in practice, of course, is that, at least as an initial matter, our parents or those who rear us must do it for us. I explore this further in Rosati (manuscript). See also Schapiro (1999).

[37] The theory of personal good that comes closest to grasping this, at least as far as I have been able to determine, is Rawls's theory of a person's good as given by her rational plan of life. But that theory suffers from a couple of fundamental difficulties. First, it appears to be vulnerable to more general objections to informed desires theories of welfare. See Velleman (1988), Sobel (1994), Loeb (1995), and Rosati (1995a and 1995b). Second, it evidently fails to recognize that being good for a person is a distinct relation that holds between an individual and an aim, undertaking, or relationship. On the nature of the *good for* relation, see Rosati (2006).

[38] Even those who are not realists about personal good would surely concede that there are more and less sensible, more and less rational, ways of organizing a life.

measure of success, or we might be unable, despite our initial desires, to cultivate an abiding interest in them. An ordering that fits our nature must suit our individual abilities, temperaments, and interests as they are and as they might develop. But in addition, they must suit us as autonomous agents – creatures who engage in self-reflection, who care what we are like and how we are motivated, who seek to define ourselves by what we choose. Not just any good will be one that we can autonomously pursue or one involvement with which will enable us, over time, to fare well and function well as the autonomous agents we are.[39] For both of these reasons, only some among the things that we may desire will be things we can successfully make good for ourselves by engaging with them in the requisite ways.[40] Only some of these things can figure in an effective conception of our good.

Still, among the many things that a person could make a part of her good, she will have to choose, and in so doing, she will shape who she is and what her life is about. She will select some ends and aims over others and, in the process, some ways of thinking, feeling, and being motivated. In ordering her desires and developing a conception of her good, a person comes at the same time to embrace an ideal of the person, a conception of the sort of person she will be; she thereby engages in a kind of self-invention. As persons, we are motivated to preserve our autonomy, and in order to function well as self-governing beings, we must order our desires in a way that gives us someone to be. By devising a conception of our good and so adopting a self-ideal, we provide ourselves with standards that guide our feelings and actions. We make ourselves people for whom some things rather than others will matter, for whom some responses rather than others will be appropriate, for whom some things rather than others will be good.

6.

Once we see our good as a solution to the practical problem we confront, we are in a position to explain the possibility of well-grounded regret,

[39] I explore further the relationship between faring well and functioning well in Rosati (manuscript).

[40] I explore briefly how one comes to make something a part of one's good in Rosati (2006). The idea that we make things a part of our good may seem odd, but the intuitive idea is in fact quite familiar. Many things come to be a part of our good only because we have taken them up and become invested in them – think, say, about the commitment involved in a marriage. Only then are we adversely affected if our relation to those things is upset.

which can take a number of forms. One basic form concerns choices that subvert our conceptions of our good. Insofar as a person gives herself an effective conception of her good, she begins to give content to the notion of the regrettable, creating correctness conditions for her choices: she thereby makes both possible and avoidable many specific regrets by providing a foundation for the feeling that she can choose wrongly. Thus, when a person experiences regret, she may indeed feel that she has chosen contrary to her good. Some of our regrets, of course, will concern desires that we will have whatever our conceptions of our good might be – desires, for instance, to be healthy and well fed. These regrets can be well-grounded whether or not we have succeeded in giving ourselves an effective conception of our good.

Of course, our regrets can run more deeply. For our conceptions of our good may be less than fully adequate, and so our regrets may reflect our failure to order our aims successfully. Perhaps we have adopted aims that are, as a practical matter, incompatible, or aims that will continually lead to disappointment. In such cases, we have indeed made a mistake about our good but a mistake of a different order. Even when they are effective, our conceptions of our good and our self-ideals typically develop and change as we refine and reinterpret them and as we change through the choices that we make, or through the changes our circumstances force on us. We may come to see inadequacies in our conceptions of our good, or we may change in ways that render them inadequate. Confronted with such difficulties, it is open to a person to reshape her self and her good. She can reinterpret what she is about and revise and reorder her aims; she can devise a more effective conception of her good, adopting a different self-ideal to guide her feelings and choices. It is partly for this reason that human lives are not the sort of thing that tends to get ruined once and for all.[41]

Finally, we may be cognizant of other lives that might have been our good and other selves we might have been, if only we had chosen them. Or we might simply be aware of the good things we missed out on because the good life we have lived precluded them. While we will not in such cases judge that we have chosen mistakenly, we may well feel that something undesirable has occurred, and in judging this, our regret is surely well-grounded. In sum, because we must devise conceptions of our good and because we can carry out this task more and less successfully,

[41] For further reasons, see Velleman (1991). Velleman argues that although certain experiences might seem to ruin a life or to permanently depress its value, a life can in fact often be redeemed by later experiences that fit the earlier ones into a narrative structure that enhances the value of the life (or an extended period within it) taken as a whole.

the possibility of well-grounded regret rests, in fundamental ways, with us.

As we make our way through life living out a conception of our good, our view of our choices and our corresponding regrets are understandably subject to change. For even our unsuccessful choices typically result in the acquisition of new desires and projects, and these may alter both our sense of who we are and our view of the unsuccessful choices themselves (Williams 1981, p. 30). We interpret and reinterpret our pasts in light of later choices and experiences, which we may regard as redeeming or damning how we spent our earlier time. As a consequence, our regrets shift over time as the significance we attach to earlier choices and experiences changes. These shifts take place not simply because of changes in our epistemic position but because our regrets and our assessments depend on our conceptions of the good and on our self-ideals. Our judgments and our regrets may thus depend upon how past choices accord with how we now conceive of our lives and who we now aspire to be. Our self-ideals, for example, can even prescribe how to react when things don't turn out as we had hoped or anticipated. How we assess our choices depending upon how they turn out is a function of luck not only because of luck in how they turn out, but because of luck in who we are, how we conceive our good, and what self-ideals we hold when assessing them at a later time. Whether our earlier choices will be regretted and assessed as in some sense undesirable depends not just on how they fare but on who we have brought ourselves to be through our choices.

To be sure, we may rule out certain interpretations of past choices as self-deceived or distorted, and we may arrive at some settled conviction about whether our choices were mistaken or yielded undesirable results. But as autonomous agents, creatures who must construct conceptions of our good and self-ideals that guide our choices and assessments, we can interpret our experiences in various ways, in accordance with the diverse goods and ideals we might adopt for ourselves.

Sometimes we are attracted to conflicting lives and self-ideals, each of which we think merits being chosen, and each of which we think we make some mistake by not choosing. This may lead us to regret and not regret a choice at the same time. When a person has such mixed feelings, of course, she may regret and regard a choice as mistaken in some respects but not others and simply be unable to decide whether she regrets it all things considered. But a person who is drawn to competing lives and selves need not always think that conflicting interpretations of her choices can be reconciled. She can interpret and reinterpret her

experiences in the process of shaping herself and her life and even choose to live with mixed interpretations, as an actor lives with mixed reviews of his performances.

The circumstances of the good thus explain not only why we feel regret and why we must develop a conception of our good but why our regrets shift as they do. Because our complex circumstances necessitate reflective choosing, regret can get a foothold. Because we must contend with our circumstances as creatures who engage in self-invention and who reflect on and reinterpret our choices, regret not only gets a foothold but undergoes its curious vicissitudes. The normal exercise of our capacities as autonomous agents thus has the peculiar result that regret remains a live possibility even with respect to some possibilities that were foreclosed without regret.

Yet our regret is not continuous. When we choose among competing lives and selves, after all, we choose among options that attract us, and where the choice of a life was not completely misguided, we are bound, in the course of living it, to find some satisfaction in it. Moreover, while we have the capacity to reflect on our lives and choices, we spend most of our time living our lives rather than reflecting on them, and that is surely the rational way to proceed. If we are to ensure that at least some of our desires are satisfied and that we lead satisfying lives, we cannot merely choose between lives: we must live the lives we choose. It is for these reasons that, although we are subject to regret constantly, we do not feel it constantly, or at least, few of us do.[42] Instead, we are more likely to feel regret when important events trigger our reflection, reopening for us the question of who to be and how to live, events such as a birth or a death, attending a school reunion, an advancement or a setback in our careers, or simply reaching a critical age (consider the common phenomenon of "mid-life" crisis). Events such as these cause us to reflect upon the possible lives and selves we have left behind and on the limited time remaining ahead, thereby bringing to the fore the circumstances of the good.

7.

I indicated earlier that I accept, suitably understood, the intuition that a good life is a life without cause for regret. If my story about our regret

[42] One further reason for this is that some of us may adopt ideals that proscribe regret, ideals that tell us to look forward, not back, to focus on our present and future, not dwell on the past.

and our good is correct, we can now see how to understand that intuition: a good life is one without cause for regret in the sense that it is one in which a person has navigated well the circumstances of the good. She has succeeded in building a life in which she both fares well and functions well. Still, as we have seen, even the person who manages to develop an effective conception of her good and to follow it out effectively may not thereby manage to avoid regret. Well-grounded regret remains a real possibility even for her.

The circumstances that explain why we feel regret thus set important limitations to the avoidability of regret. Because we live our sole and limited lives as persons, we will change and reflect upon and reevaluate our choices, while living in awareness of the significance of our choosing for who we are and for the character of our short lives. When I say that regret is, in an important sense, unavoidable, I do not mean to exaggerate the extent to which our lives must be regret-ridden. While some people may feel regret intensely and frequently, as I have explained, most of us will feel it more sporadically, and some of us hardly at all. My point is that regret is unavoidable for deep reasons, and not simply because we lack knowledge or feel pain or because we believe that we can fail to fasten upon our best lives. Its unavoidability reflects the condition of creatures whose circumstances drive them to invent themselves and their good.

Acknowledgements

For helpful discussion and comments on various versions of this paper, I want to thank David Anderson, Sarah Brody, Sarah Buss, Josh Cohen, Stephen Darwall, Richard Dees, John Deigh, Julia Driver, Don Herzog, Allan Gibbard, Elijah Milgram, Richard Moran, and David Velleman. Many thanks also to Jonathan Adler, Jeff Blustin, Chris Gowans, and Michael Stocker, who provided helpful feedback during a meeting of their discussion group, and to members of the philosophy departments at the University of North Carolina, Cornell University, the University of California, Davis, and the College of William and Mary.

University of Arizona
Department of Philosophy
Social Science Building, Room 213
Tucson, AZ 85721
e-mail: csrosati@email.arizona.edu

REFERENCES

Aristotle (1926). *The Nicomachean Ethics*. Translated by D. Ross. Oxford: Oxford University Press.

Baron, M. (1988). Remorse and Agent-Regret. In: P.A. French, T.E. Ueling, Jr., and H.K. Wettstein (eds.), *Midwest Studies in Philosophy*, vol. XIII, pp. 259-281. Notre Dame, IN: University of Notre Dame Press.

Bittner, R. (1992). Is It Reasonable to Regret Things One Did? *The Journal of Philosophy* **89**, 262-273.

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

Bratman, M.E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Brink, D.O. (1987). Rawlsian Constructivism in Moral Theory. *Canadian Journal of Philosophy* **17**, 71-90.

Buss, S. (1994). Autonomy Reconsidered. In: P.A. French, T.E. Ueling, Jr., and H.K. Wettstein (eds.), *Midwest Studies in Philosophy*, vol. XIX, pp. 95-121. Notre Dame, IN: University of Notre Dame Press.

Cullity, G. and B. Gaut, eds. (1997). *Ethics and Practical Reason*. Oxford: Clarendon Press.

Darwall, S.L. (1983). *Impartial Reason*. Ithaca, NY: Cornell University Press.

De Sousa, R.B. (1974). The Good and the True. *Mind* **83**, 534-551.

Dworkin, G. (1970). Acting Freely. *Nous* **4**, 367-383.

Frankfurt, H.G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* **68**, 5-20.

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

Grice, P. (1991). *The Conception of Value*. Oxford: Clarendon Press.

Griffin, J. (1986). *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.

Kant, I. (1964). *Groundwork of the Metaphysic of Morals*. Translated by H.J. Paton. New York: Harper and Row.

Kolnai, A. (1962). Deliberation Is of Ends. *Proceedings of the Aristotelian Society* **62**, 195-218.

Korsgaard, C.M. (1983). Two Distinctions in Goodness. *Philosophical Review* **92**, 169-195.

Korsgaard, C.M. (1989). Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy and Public Affairs* **18**, 101-132.

Korsgaard, C.M. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

Loeb, D. (1995). Full-Information Theories of Individual Good. *Social Theory and Practice* **21**, 1-30.

Nagel, T. (1979). Moral Luck. In: *Mortal Questions*, pp. 24-38. Cambridge: Cambridge University Press.
Nagel, T. (1986). *The View from Nowhere*. New York: Oxford University Press.
Neely, W. (1974). Freedom and Desire. *Philosophical Review* **83**, 32-54.
Parfit, D. (1973). Later Selves and Moral Principles. In: A. Montefiore (ed.), *Philosophy and Personal Relations*, pp. 137-169. London: Routledge and Kegan Paul.
Parfit, D. (1975). Personal Identity. In: J. Perry (ed.), *Personal Identity*, pp. 199-223. Berkeley, CA: University of California Press.
Parfit, D. (1984). *Reasons and Persons*. Oxford: Clarendon Press.
Railton, P. (1986a). Moral Realism. *Philosophical Review* **95**, 163-207.
Railton, P. (1986b). Facts and Values. *Philosophical Topics* **14**, 5-31.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, J. (1980). Kantian Constructivism in Moral Theory: The Dewey Lectures 1980. *The Journal of Philosophy* **67**, 515-572.
Rawls, J. (1993). *Political Liberalism*. New York: Columbia University Press.
Raz, J. (1986). *The Morality of Freedom*. Oxford: Clarendon Press.
Rorty, A. (1980). Agent-Regret. In: A.O. Rorty (ed.), *Explaining Emotions*, pp. 489-506. Berkeley: University of California Press.
Rosati, C.S. (1995a). Naturalism, Normativity, and the Open Question Argument. *Nous* **29**, 46-70.
Rosati, C.S. (1995b). Persons, Perspective, and Full Information Accounts of the Good. *Ethics* **105**, 296-325.
Rosati, C.S. (2000). Brandt's Notion of Therapeutic Agency. *Ethics* **110**, 780-811.
Rosati, C.S. (2006). Personal Good. In: M. Timmons and T. Horgan (eds.), *Metaethics After Moore*, pp. 107-132. Oxford: Oxford University Press.
Rosati, C.S. (manuscript). *Autonomy and Personal Good: Lessons From Frankenstein's Monster*.
Sayre-McCord, G. (1986). The Many Moral Realisms. In: G. Sayre-McCord (ed.), *Essays in Moral Realism*, pp. 1-23. Ithaca, NY: Cornell University Press.
Scanlon, T. (1982). Contractualism and Utilitarianism. In: A. Sen and B. Williams (eds.), *Utilitarianism and Beyond*, pp. 103-128. Cambridge: Cambridge University Press.
Schapiro, T. (1999). What Is a Child? *Ethics* **109**, 715-738.
Smith, M. (1997). A Theory of Freedom and Responsibility. In: Cullity *et al.* (1997), pp. 293-320.
Sobel, D. (1994). Full Information Accounts of Well-Being. *Ethics* **104**, 784-810.
Spinoza, B. (1951). *Ethics*. Translated by R.H.M. Elwes. New York: Dover.
Stampe, D, (1987). The Authority of Desire. *Philosophical Review* **96**, 355-381
Stocker, M. (1979). Desiring the Bad: An Essay in Moral Psychology. *The Journal of Philosophy* **76**, 738-753.
Taylor, C. (1985a). What Is Human Agency? In: Taylor (1985c), pp. 15-44.
Taylor, C. (1985b). Self-Interpreting Animals. In: Taylor (1985c), pp. 45-76.
Taylor, C. (1985c). *Human Agency and Language: Philosophical Papers*, vol. I. Cambridge: Cambridge University Press.
Velleman, J.D. (1988). Brandt's Definition of 'Good'. *Philosophical Review* **97**, 353-371.
Velleman, J.D. (1989). *Practical Reflection*. Princeton, NJ: Princeton University Press.
Velleman, J.D. (1991). Well-Being and Time. *Pacific Philosophical Quarterly* **72**, 48-77.
Velleman, J.D. (1992). The Guise of the Good. *Nous* **26**, 3-26.
Velleman, J.D. (1996). The Possibility of Practical Reason. *Ethics* **106**, 694-726.
Velleman, J.D. (1997). Deciding How to Decide. In: Cullity *et al.* (1997), pp. 29-52.
Watson, G. (1975). Free Agency. *The Journal of Philosophy* **72**, 205-220.

Williams, B. (1973a). The Makropulos Case: Reflections on the Tedium of Immortality. In: Williams (1973c), pp. 82-100.

Williams, B. (1973b). Ethical Consistency. In: Williams (1973c), pp. 166-186.

Williams, B. (1973c). *Problems of the Self*. Cambridge: Cambridge University Press.

Williams, B. (1981). Moral Luck. In: *Moral Luck: Philosophical Papers 1973-1980*, pp. 20-39. Cambridge: Cambridge University Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Translated by G.E.M. Anscombe. New York: Macmillan.

Zoch, L.N. (1986). Remorse and Regret: A Reply to Phillips and Price. *Analysis* **46**, 54-57.

G.F. Schueler

RATIONALITY AND CHARACTER TRAITS

We frequently explain human actions, or at least try to explain them, by citing the agent's reasons for doing what she did. And a great deal of work in contemporary moral psychology has gone into attempts to explicate how such purported explanations of action do their job, that is, how they succeed in explaining actions. Much of this effort of course has focused on the so-called "belief-desire" account of agents' reasons. But there is a problem here. It is not obvious that the agent's reasons for doing whatever she did are actually enough to explain her action. Here is how Thomas Nagel puts this problem.

> When someone makes an autonomous choice such as whether to accept a job, and there are reasons on both sides of the issue, we are supposed to be able to explain what he did by pointing to his reasons for accepting it. But we could equally have explained his refusing the job, if he had refused, by referring to the reasons on the other side – and he could have refused for those other reasons: that is the essential claim of autonomy. It applies even if one choice is significantly more reasonable than the other. Bad reasons are reasons too. Intentional explanation, if there is such a thing, can explain either choice in terms of the appropriate reasons, since either choice would be intelligible if it occurred. But for this very reason it cannot explain why the person accepted the job for the reasons in favor instead of refusing it for the reasons against. (Nagel 1986, pp. 115-116)

Clearly it will not be enough here merely to say that the one set of reasons, instantiated so to speak in one set of the agent's desires and beliefs, *caused* him to take the job. That simply raises the same problem in slightly different terms. If he had refused to take the job we could equally "explain" his refusal by saying that the other set of desires and beliefs had caused him to refuse it. The same question Nagel raises would then arise again in causal terms: why did one set of desires and

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 261-278. Amsterdam/New York, NY: Rodopi, 2007.

beliefs become causally efficacious while the other set did not. So there seems to be a real issue here.

In what follows I want to look at some reasons for thinking that action explanations of the sort Nagel is discussing need to involve a reference to one or more traits of character of the agent whose action is being explained. This is true I think even in the most straightforward cases of explanation in terms of the agent's reasons. At the same time, if this is true, it suggests a way of solving the puzzle that Nagel raises. We may begin by considering a different example.

Let's suppose that in a local election I am considering whether to vote for a tax increase the income from which is earmarked toward subsidizing day care in my community. I reason as follows:

(1) "Subsidized day care is a good thing," I say to myself.
(2) "This proposed tax increase is necessary if there is to be subsidized day care in my community."
(3) "At the same time, it will cost me some money, which I would like to use elsewhere, if this tax increase is passed."
(4) "Still, it is more important that my community have subsidized day care than that I keep for my own use the few bucks it will cost me each year."

"So, I should vote for this tax increase."

On the basis of this reasoning, let's suppose, I do indeed vote for the tax increase. At the same time my neighbor, who votes against this tax increase, accepts word for word the first three premises here (again, let's suppose) but reverses the weight she assigns each of the two facts (or whatever they are) described in the first and third premises. So the forth premise she accepts is not the one above but

(4′) "It is more important that I keep for my own use the few bucks it will cost me each year than that my community have subsidized day care."

What explains the difference in our reasoning (and hence the difference in our votes)? One possibility of course is that our circumstances are different. It could be that her need for those "few bucks" the tax increase will cost each of us is very much greater than mine. So it could be that, given the difference in our circumstances, each of us reasoned correctly. (This raises the question of whether for this to happen the two different fourth premises, (4) and (4′), wouldn't need to be relativized to the two different agents. Does "more important" here

mean "more important for me" or "more important, period" so to speak? Let's just set this issue aside for the moment. We will return to it below.)

But let's suppose that our circumstances are relevantly similar, with incomes, expenses, and so forth that don't differ importantly. And to keep it straightforward let's also suppose that neither of us misjudges our circumstances or how much extra tax we will have to pay. In short, let's suppose that in our life circumstances, moral views (other than that contested fourth premise) and knowledge of whatever we regard as relevant to this issue, the two of us are essentially the same.[1] The only place we differ is about how to weigh the public good of subsidized day care against our own private interest, that is, on the forth premise of two pieces of reasoning described above.

In this circumstance it would seem that the obvious explanation for the difference in our reasoning, and hence the difference in our votes, is that we are somewhat different kinds of people, with different values or priorities, that is that, really, we have somewhat different characters. I am *the sort of person* who is willing to put aside my own interest, at least when it is a relatively minor one like this, in favor of what I think is a public good (sterling fellow that I am), while my neighbor is not. That is, what this sort of comparison between my neighbor's vote and mine suggests is that to understand either one of us we must refer to more than just the considerations each of us acts on. We must also bring in a reference to the kinds of people we are, people who evaluate the same considerations in different ways.

One reason it is easy to miss this point is that the usual account of the agent's reasons as consisting of desires and beliefs invites the idea that this reasoning can be adequately understood in terms of the traditional practical syllogism. As Colin McGinn puts it, "Whenever someone acts for a reason we can assume some such reasoning [as is represented in the practical syllogism] to have occurred. We can thus say that an action is a bodily movement issuing from such practical reasoning as is codified in the practical syllogism" (McGinn 1979, p. 24)[2]. The problem is that this simply ignores the countervailing considerations that agents typically consider when they deliberate. Instead it focuses only on the single factor

[1] All this agreement is intended to rule out, along with much else, any "ideological" differences about whether it would be better to have only privately funded day care rather than publicly funded facilities, for instance. More importantly though, this example is intended to be a "two-person" analogue of the "one-person" case Nagel describes. So of course differences have to be kept to a minimum. This should become clearer below.

[2] As McGinn makes clear, this is of course not to say that agents always engage in explicit reasoning whenever they act. They obviously do not.

that the agent in question ends up regarding as the one to act on. So on this sort of account my reasoning in the case we are considering gets transfigured into something like this:

(1) "Subsidized day care is a good thing," I say to myself.
(2) "This proposed tax increase is necessary if there is to be subsidized day care in my community."

"So, I should vote for this tax increase."

By the same token, my neighbor's reasoning gets represented as:

(3) "It will cost me some money, which I would like to use elsewhere, if this tax increase is passed."
(3.5) "The only thing I can do toward stopping this tax increase is to vote against it."

"So, I should vote against this tax increase."

If taken literally either as accounts of the reasoning my neighbor and I *actually* used or as accounts of the reasoning we *should* have used, both of these two syllogisms are simply inaccurate given the way I have described things here. Each leaves out both any reference to the consideration on which we did not act and any description of the weight or evaluation each of us gave the two things we considered (which I am representing in the two different versions of the fourth premise in the two pieces of reasoning described above). Understanding practical reasoning in terms of the practical syllogism thus can easily lead one to ignore completely the possibility that two agents can each be aware of exactly the same set of considerations, moral and factual, and yet, from differences in character, weigh these considerations differently (a situation which in ordinary life seems not uncommon).

One might say that except for the simplest of cases, the practical syllogism only accurately describes the practical reasoning of the very narrow-minded or the very badly informed. But even this is too generous since practical reasoning only occurs, only makes sense in fact, where a choice is possible, and the practical syllogism in its traditional form is incapable of giving any account of the weighing up of pros and cons that is the essence of deliberating about how to choose.[3]

A second, perhaps less obvious, reason it is easy to miss the importance of referring to the agent's character in explaining her actions is that it is easy to make the mistake of understanding either my or my

[3] Davidson discusses this defect in the practical syllogism in his (1980).

neighbor's version of the fourth premise in our reasoning as referring to "what is important *to me*" rather than simply to "what is important," period. Of course the fact that we accept these premises, and use them in our reasoning, shows (to some extent) what is important to us. But it would be a mistake to read this fact into the *content* of the reasoning that my neighbor and I use.

Someone who weighs the public good of subsidized day care against her own interest in retaining the few extra dollars this would cost her and then decides that the former is more important (or that the latter is) is *already* thereby weighing her own concerns against those of the public at large. The version of the fourth premise that I accept, or rather the fact that I accept and act on it, shows that I am the sort of person who regards the public interest in this situation as more important than my own interest in those extra few dollars, *not* that I am the sort of person who regard's the public interest as more important *to me*.

That would be a different thought, a *factual* belief about the sort of person I am (or about what I regard as important) rather than an *evaluative* belief about the relative importance of subsidized day care and my having a few extra bucks each year. The public interest *is* more important to me (at least in this case). That is shown by the fact that I accept and act on (4) above. People of course sometimes hold that their own concerns are more important than those of the public generally, that is how I am portraying my neighbor in this example. But to understand the two people who reason in terms of the two versions of the fourth premise above as necessarily referring (at least implicitly) to what is important "to them" is to misstate the difference between my neighbor and myself. It would be to mistake what might be called an external or observer's description of our attitudes for the *content* of our reasoning. An external observer would be quite right to say that having subsidized day care is more important to me, that is, something I want more, than having those few extra bucks, and vice versa for my neighbor. But what makes this description true, in my case, is that I think our community having subsidized day care is more important than my having those few extra bucks, and that thought is what moves me to vote as I do.

It is, perhaps, clear enough how understanding practical reasoning simply in terms of the practical syllogism (that is, simply in terms of the primary desire-belief pair on which the agent acts) covers up the importance of referring to the character of the person moved by the reasoning. In the case we are discussing it does this by ignoring the need to refer to different versions of the fourth premises in the reasoning my neighbor and I engage in. But the second sort of mistake just described

also covers up the importance of referring to the characters of the those doing the reasoning by making it seem as if both my neighbor and I only look at *our own* concerns in deciding how to vote. Reading "importance" in the two versions of the fourth premises of our reasoning as "importance for me" makes it seem as if all practical reasoning is necessarily egocentric or self centered, in the sense that the only thing anyone considers, or even can possibly consider, is her own interests and concerns.

Speaking of differences in character in such a situation will then seem disingenuous. Everyone will appear as profoundly, indeed *necessarily*, focused only on his or her own concerns. But in fact, as I hope is clear, this is just a mistake. Neither of us, in the reasoning I described, looks *only* at our own concerns. Each of us weighs, that is evaluates the importance of, our own concerns, in this case represented by premise (3), against the public good of subsidized day care, which each of us also evaluates positively.

To understand the two versions of the fourth premise in this reasoning as referring for each agent to what is important *to her* would be to make both the original two pieces of reasoning described above automatically and obviously fallacious. In fact, if accepted generally, it would seem to make any sort of non-fallacious practical reasoning impossible. As stated above the premises of the original two arguments at least appear to support the conclusions drawn (and do in fact support it, I want to say). But if "more important" is understood as "more important for me" in each case, each piece of reasoning becomes an example of the fallacy of *ignoratio elenchi*.[4] This is because the fourth premise in each piece of reasoning would now no longer evaluate the importance of the first and third considerations, and hence if true provide reason to accept the conclusion. Each version of the fourth premise would now simply report a fact about the person doing the reasoning and as such would become completely irrelevant to the question of which way she *should* vote.

Genuinely evaluative premises, of the sort represented by the two versions of the fourth premise in the two pieces of practical reasoning we have been looking at, are essential to practical reasoning. (Or at least so I claim.) Accounts of practical reasoning that leave such premises out, as the standard "practical syllogism" account does according to some ways of understanding it,[5] or which distort evaluations into factual claims about the attitudes or psychology of the person doing the reasoning, as reading "important" to mean "important to me" in those evaluative

[4] This is the fallacy also known (more prosaically) as "irrelevant conclusion."

[5] See for example Robert Audi's discussion of the practical syllogism in his (1989, p. 99).

premises would do, threaten to make genuine, non fallacious practical rationality impossible.[6] And they give a false account of the agent's reasons. They make it seem, one might say, that even the agents themselves are not so much *doing* the reasoning as observing or describing themselves doing it.

It is possible, of course, to describe the facts about the characters of my neighbor and myself that are revealed in our acceptance of the two different forth premises above as facts about different pro attitudes we are thereby revealed to have, or perhaps in terms of different strengths of pro attitudes toward our own interests and subsidized day care. By itself there is nothing objectionable in this unless we mistakenly think it is the whole story. To do that would be to ignore how my neighbor and I understand our own actions. That is, what seem to me and to my neighbor to be actions *based on* reasoning involving different judgments about the relative importance of a public good and our own interests would be interpreted in terms of "pro attitudes" and then explained as being caused by different pro attitudes, perhaps of different strengths, toward the public good and self-interest. But this simply gets the explanatory story backwards. In the case we are examining the only evidence an observer could have that I *have* a pro attitude toward publicly subsidized day care is that my belief that it is a good thing is used by me in the reasoning described above as a reason for voting for that tax increase.

So reference to both the evaluative beliefs and the character traits of the person acting on the basis of this reasoning is simply not eliminable in favor of some set of pro attitudes and beliefs. This is one of the reasons the sort of explanation of action I am describing is essentially a personal level explanation.[7] References to character traits as elements of explanations of actions will seem deeply problematic, of course, if one thinks of them the way Ryle for instance often seems to, as nothing but "propensities," that is if one thinks of the names of character traits as standing for nothing but dispositions to behave in certain ways (Ryle 1949).

It is hard to see how dispositions of this sort could be used to *explain* the behavior at issue in any but the "virtus dormativa" sense of "explain."

---

[6] They thus perhaps help clear the way for "sub-personal," or causal, accounts of the explanatory force of reason explanations since they make it seem that the reasoning agents use to decide what to do is never really any good. See note 7 below.

[7] Jennifer Hornsby describes the personal level as "a level at which mention of persons is essential, and [. . .] commonsense psychological explanations are indigenous to that level" (Hornsby 1997, p. 161). Hornsby has an excellent discussion of some of the ramifications of the personal-subpersonal distinction. The distinction itself comes from Daniel Dennett (see Dennett 1989).

On this account of character traits we would "explain" the fact that someone acted impulsively, say, by referring to the fact that he was "disposed to be impulsive," which of course doesn't explain his impulsiveness at all. This problem seems to force anyone who wants to include character traits in the account of explanations of actions into giving an analysis of these traits in terms of their underlying features, perhaps on the model of the account in terms of molecular structure of the dispositional property some objects have of being water-soluble. And what more obvious underlying features for analyzing character traits than some sets of pro attitudes? But I want to argue that neither the "plain" (Ryle-like) "dispositional" account nor any "set of pro attitudes" account of character traits gets them right. This will perhaps be clearer if we shift to a different case and a different character trait.

Suppose I know a generous person. By seeing that this person has this character trait I am able to understand, give an explanation of, why, in certain circumstances, she acts and reasons as she does. Of course for any particular action in which this character trait plays a role I will be able, if I know enough of the details, to describe her action in terms of her factual beliefs and "pro attitudes" (some of which may just be evaluative beliefs that figure into practical reasoning on the basis of which she acts). But knowing that she is generous I will also be able to see *why* she has some of these evaluative beliefs and other pro attitudes.

It might seem possible that we could if we worked at it spell out what it is to be generous in terms of some set of pro attitudes. But even if we managed to identify a set of attitudes that as a matter of fact fit all and only people who were generous (above some minimum), it would be a mistake to think that we had somehow done away with the need to refer to generosity *itself*, the character trait. People after all sometimes do generous things, and thus have the relevant particular pro attitude, but don't do them *from generosity*. And that difference needs to be accounted for. There is a difference between someone who does something generous, say gives some of her money to someone who needs it, on a whim or from a momentary feeling of generosity or to impress a friend or salve her conscience, and someone who actually gives this money *out of generosity*.[8]

[8] There might be cases where we were inclined to say that someone had acted "out of generosity" even though she wasn't "a generous person," that is, where she did not have the character trait of generosity. She acts "out of character" as we say. I suspect that such cases are ones where the agent acts from "a feeling of generosity" but whether or not that is right, this and numerous other cases raise the issue of the exact nature of the

So even if we could set out completely the set of beliefs and attitudes that are in some sense extensionally equivalent to the character trait of generosity as it actually occurs, we would lose something in explaining some act only by reference to some members of this set of attitudes and beliefs. We would still need to know why this particular set of beliefs and attitudes hang together, as a unit, so to speak, and how that unit *itself* figured into explaining the action in question (when it did), before we could get from explaining some particular action on the basis of one or a few of these attitudes and beliefs to seeing it in terms of the whole set, that is in terms of the character trait.

Suppose I discover that you gave someone your old car. I find out something of why you did this when I find out that you wanted, or perhaps thought it important, to do something to help him out and thought that this would do that. (As opposed to wanting to get that eyesore out of your driveway, for instance.) But your pro attitude toward doing something that would help him out, by itself, connects to nothing further, even if I know that you also have the other attitudes and beliefs in the extensionally described set supposed to be coextensive with generosity. The mere fact that you *have* all these other attitudes and beliefs in this set, by itself, does nothing to explain this particular act of yours. All those other pro attitudes and beliefs in this set are, so far at least, simply *idle* as far as this explanation is concerned. There is still the question: Why did you want (or think it important) to help him out? And the mere fact that you have all those other attitudes and beliefs supposed to constitute, or be coextensive with generosity, by itself, does nothing to answer that question.

There may of course be a further attitude or belief, a further reason of yours, that explains the one on which you acted: perhaps you wanted him to be in your debt. If that really was your reason for wanting to help him out then your action doesn't look quite so generous after all. That particular attitude, desire to put someone in your debt, probably isn't a member of the set supposed to be coextensive with the character trait of generosity. But whether or not there is a further reason of yours to explain why you wanted to help him out, any such explanation in terms of your reasons will come to an end somewhere and I want to say that one such place is in a character trait such as generosity. It could be for instance that you don't have any further reason, such as a desire to put him in your debt, to explain why you wanted to help him out. It could be

---

explanation being offered when a reference is made to character traits in these ways. This issue will be taken up again below.

that you wanted to help him out simply because you are a generous person.

There are two points here. Reference to a character trait such as generosity *explains* (or certainly seems to explain) the having of certain beliefs and attitudes in a way that merely citing the fact that the attitude in question is a member of some extensionally described set cannot. At the same time reference to a character trait moves outside the "chain of reasons" and so doesn't leave the explanation hanging, so to speak, in the way explaining one attitude in terms of another that constitutes your reason for having it does. If we find out that you wanted to help him out because you wanted to put him in your debt, we are left with the question of *why* you wanted to put him in your debt. That is, we are left with the same *sort* of question all over again, in a way that we are not when we find that you wanted to help him out because you are a generous person.

The difference is that attitudes such as wanting to help someone out and wanting to put someone in your debt are things that can themselves be explained in terms of the agent's reasons for having them. So when we explain your wanting to help him out in terms of your wanting to put him in your debt we are explaining one thing for which you can have a reason, your wanting to help him out, by citing another thing for which you can have a reason, your wanting put him in your debt. That clearly leaves us with the question of what your reason was for having that second attitude, wanting to put him in your debt. And this *sort* of question will arise whether or not the second attitude is a member of the appropriate extensionally described set. Nothing about the fact that the second attitude is (or is not) a member of that set means that you couldn't have it "for a reason" as well.

But character traits are typically not things which one has for a reason, at least in anything like the same way. Though you may well have a reason for wanting to put someone into your debt, and then give him your old car for that reason, it doesn't seem likely that you are a generous person "for a reason." Of course the fact that you are generous, like any fact, may have an explanation, and in *that* sense we may speak of "a reason why you are generous" (just as there may be a reason why you are nearsighted or right handed). Your generosity might be explained by the way you were raised, the sort of parents you had or the like. But if generosity is a character trait then it is not something like giving someone your old car or wanting to put him in your debt, where it makes sense to speak of your reason for doing or wanting these things. The shift to character traits in this way is so to speak a shift out of the first person

point of view, of the sort required when referring explicitly to "the agent's reasons," to an essentially external or observer's point of view.[9]

So when we explain your wanting to help him out by citing the fact that you are a generous person, the explanation in terms of your reasons comes to an end at that point. The claim I am making then is that at least one clear way explanations of actions in terms of the agent's reasons can come to a satisfactory end is by referring, either explicitly or implicitly, to some character trait. This is a satisfactory end in the sense that we are not left with a question of exactly the same sort still unanswered, as we are if the answer is in terms of an attitude which is itself one you can hold for a reason.

In short, when we explain your giving him your old car by referring to the fact that you want to help him out, one of two things happens. Either you have a reason for wanting to help him out, a reason we need to find to understand your action. Or in finding this particular attitude of yours we have found out what *sort* of person you are, the sort who wants to help someone in this kind of circumstance. If we understand the answer in this second way then the explanation in terms of your reasons comes to an end, no further explanation is needed or appropriate, since we have found a feature of your character which itself explains the attitude. It is very much the same kind of thing that happens when we find that someone is doing something in order to avoid pain. It is not that one couldn't have a reason for wanting to avoid pain, since of course one could. It is simply that virtually all normal humans are the sorts of people (or the "sorts of beings") who do indeed want this.[10]

It may be worth noting in passing that the account of the place of character traits in explanations of actions in terms of the agent's reasons that I am suggesting seems not to be subject to the objection on which Ryle's more minimalist account of character traits seems to founder. As was mentioned above, Ryle often speaks as if character traits are *mere*

---

[9] Typically at least. The issue of whether one can have such a trait intentionally will be taken up below.

[10] "Ask a man *why he uses exercise*," Hume says, "he will answer *because he desires to keep his health*. If you then enquire, *why he desires health*, he will reply, *because sickness is painful*. If you push your enquiries further and desire a reason *why he hates pain*, it is impossible he can ever give any. This is an ultimate end and is never referred to any other object" (Hume 1975, p. 293).

One might think Hume is not quite right here since of course one can have a reason for wanting to avoid pain, e.g., it could be distracting. But in fact he speaks of "hating pain" and in that I think he is right. It is a feature of humans, or normal ones anyway, that they "hate pain." We might not call this feature a "character trait" but I think that this is only because, unlike generosity, say, or selfishness, hating pain seems virtually universal.

dispositions, while at the same time he disallows any account of these traits in terms of their underlying features (Ryle 1949).[11] That seems to leave character traits as "dormitive virtues" of the sort Moliere joked about. It would mean "explaining" a bit of generous behavior by reference to the character trait of generosity and at the same time saying that generosity was nothing but the propensity to do generous things.

Since I have argued, more or less following Ryle, that character traits are not reducible to sets of desires and beliefs (which could then perhaps be understood to interact causally to produce actions), it may be useful to say explicitly why the account sketched here is not subject to this same, *virtus dormitiva* objection. The central point is that on the view we have been examining character traits are essentially features of reasons explanations of actions.[12] (They are thus essentially personal level facts.) To say that someone has a certain character trait is to say, roughly, that certain sorts of facts count for her as reasons (perhaps of a certain strength) to do certain sorts of things. That is why reference to a character trait can serve to stop the "chain of reasons" by citing a feature of the agent which both explains why she took something as a reason for acting and is not itself something for which she can have a reason. No doubt there is a dispositional element in this but it is at least not the sort of uninformative circle Ryle's purely dispositional account of character traits seems to give us.

It is part of my argument here that character traits are not the sorts of things which agents have "for reasons" in the way actions or some attitudes are. But couldn't someone be a generous person intentionally, so to speak? For instance couldn't someone become convinced to be generous by reading moral philosophy, say, and finding the reasons for being generous persuasive? In one sense the answer seems to be "yes." One could be convinced to adopt what we might call a policy of generosity, of doing the generous thing. And I think there is no doubt that someone who followed such a policy, even with the occasional lapse, would be thought to "be generous" or to "be a generous person."

To see what is happening here it might help to contrast the kind of case we have been discussing, where a reasons explanation comes to an end in a reference to the agent's character, to the superficially similar but

[11] Dennett points out that we could agree with Ryle about the meanings of character trait terms while still seeking an empirical theory to explain how these traits could work as they do (Dennett 1989).

[12] Or at least the ones I want to talk about are. If, say, clumsiness is a character trait then it is not one of the ones I am referring to since it seems to apply only to things done unintentionally or accidentally and so not to things for which one could have reasons.

in fact very different case where the agent takes *the fact that* it would exemplify a certain character trait as her reason for performing some action. This is the case if for instance your reason for wanting to help him out was that helping him out would be the generous thing to do. In this sort of case the "chain of reasons" hasn't come to an end with your desire to help him out. You have a further reason for wanting this, that this would be the generous thing to do. So citing this fact as your reason still leaves us with the further question of why you want to do what is generous (why you think the fact that it would be the generous thing to do gives you a reason to do it).

The answer to that further question *might* be simply that you are a generous person in the way we have been discussing, since one way of exemplifying the character trait of generosity is consciously to try to be generous, that is, to have generosity as an explicit ideal. If that is what is happening then, again, we have arrived at a reference to a character trait and no further "reasons" question arises. But it is also possible that you have a still further reason for wanting to do what is generous, i.e. that it is not that you want to do this just because you are a generous person. A generous person is someone who *has* that particular character trait, not merely, and not necessarily, someone who does things because the acts *exemplify* generosity.[13] As was said, you might want to do something that exemplifies generosity (and so want to help him out) because you have been convinced to adopt this policy by your reading of moral philosophy. But equally you could simply want others to think well of you, i.e. because you are a self-promoter, which is a very different sort of character trait than generosity.

Someone who takes it as a reason for wanting to help someone else out that it would be the generous thing to do (or the like) might have, in Bernard Williams' phrase, "one thought too many," since this is consistent with merely wanting to appear to be generous and so consistent with not actually being generous. But in any case the point to notice is that in this sort of case, unlike the case where reference is made to the fact that the agent really is generous, the agent's reasons don't come to an end with a desire to help him out. We are still, or might be, left with the question of what reason the agent has for wanting to do the generous thing. What this means is that in this sort of case the puzzle with which this paper began will not (yet) have been solved.

To see what is happening here let's set this problem aside for a moment and return to Nagel's puzzle. If the features of character traits

[13] As Bernard Williams points out. See Williams (1985, p. 10).

we have been examining are on the right track then I think there is an interesting consequence for that puzzle. There appears to be a feature of at least some reasons explanations, their dependence on references to traits of character of the agent who performed the action in question, that seems in no way reducible to the sorts of desires and beliefs employed in reasons explanations of actions. At a minimum there seem to be at least some cases of explanation of actions in terms of the agent's reasons where reference to the agent's character is an essential part of the explanation, even if this reference is often only implicit. If, as I have argued, character traits cannot be completely explained in terms of sets of desires and beliefs, then such attitudes will not be the only elements in explanations of actions in terms of the agents' reasons. Character traits, as a distinct kind of element, will be needed as well.

Such a feature of reasons explanations raises the possibility that some such explanations are *inherently* evaluative, as the generosity example suggests. Character descriptions at least frequently involve what Williams calls "thick moral concepts" (Williams 1985, p. 129), meaning roughly that while they are required for the accurate description of someone, they also contain an ineliminable moral or evaluative element. So if the argument above is correct it follows that explanations of actions in terms of the agent's reasons at least sometimes must involve *evaluations* of a certain sort, for instance in at least some of those cases where to be complete the explanation must involve reference to some such character trait of the agent. Some character traits, such as impulsiveness, may perhaps carry no automatic evaluation, but many, such as generosity, certainly seem to.

At the same time, as was said, the considerations about character traits that we have been surveying here go some way toward solving Nagel's puzzle, the puzzle about how reasons explanations of actions actually *explain*, at least for the cases to which they apply. This is a puzzle that looks insoluble on the usual desire-belief account that underlies the traditional version of the practical syllogism. If I am right in thinking that reasons explanations of actions at least sometimes involve an implicit reference to some character trait of the agent though, then we seem to have a possible solution to Nagel's puzzle, in general at least, though specific cases may still present difficulties. The solution is to notice that merely referring to the agent's reason or reasons for doing what she did is to leave out an essential element of "intentional explanations" as Nagel calls them. That element is the relevant character trait of the agent in question. The two parallel examples above of my neighbor, who voted against that proposed tax increase, and me, who

voted for it, contain all the same elements as Nagel's example of the person who takes the job for the reasons in favor rather than refuses it for the reasons against. The only difference is that in place of Nagel's job seeker who makes one choice and not the other, I have imagined two similar people who nevertheless choose to vote differently. But the explanation in each kind of case is the same (or of the same sort). In each case the two alternative choices make sense once we notice the *sort* of person involved. By the same token, not knowing the relevant character trait or traits leaves us with just the sort of incomplete explanation Nagel presents.

Nagel's example about choosing whether or not to take a job *looks* puzzling because no details at all are given about how the agent weighed the competing reasons for and against taking the job. Once such details are filled in, even in the rather minimal way they were in the parallel voting cases, reference to the relevant character trait of the agent can dissolve the puzzle. Suppose for instance, to add such details to Nagel's case, that the reasons against taking the job were that it involved moving out of one's hometown and accepting a lower salary while the reasons in favor were that it involved lots of "on air" TV time in a big city. (It is a TV broadcasting job, say.) Just by themselves these facts still seem to leave us with the puzzle Nagel describes. But if we also know that the agent making this choice is personally vain, the puzzle dissolves.

Of course it seems unlikely that reference to a character trait will provide a *mechanical* explanation of how the reasons cited led to the action at issue, an explanation that is of the sort that seems to be required by advocates of a purely causal analysis of how desire-belief reasons explain.[14] If character traits enter into reasons explanations in the way I am suggesting, it may just be a mistake to think that the incompleteness of action explanations of the sort Nagel describes even can be, let alone must be, completed by reference to mechanical causation.

Even if the suggestion made here is accepted though, plenty of issues remain. For instance, there is the question of what happens when the agent has several different character traits and all are relevant to the act in question. Being generous does not *exclude* also being a self-promoter, for example. And more generally, human beings standardly (perhaps even necessarily) have numerous character traits, some of which clash or conflict at least some of the time. Then too, people sometimes act "out of character." Generous people do not always act generously and people who are not generous sometimes do. But though these and other problems

[14] Such causal accounts have deep difficulties of their own in any case. See, for instance, Kim (1993).

will undoubtedly be very common and will raise questions of how explanation in such cases works, the existence of such cases doesn't (or at least needn't) call into question the sort of answer being suggested here for Nagel's puzzle. That puzzle after all did not depend on any such complications. It applies to even the simplest cases of "intentional explanation" and if unsolved would just stop *any* such explanation of action in its tracks. The complexities mentioned here (and no doubt there are many others) only show that the problem Nagel describes for intentional explanation is not the only one. They remain even if the suggestion explored here for solving this puzzle, for at least some of the simplest cases, is accepted.

A deeper problem might seem to be suggested by the issue left hanging a few paragraphs above. If someone can be generous *for a reason* then how can appeal to a character trait such as generosity serve to halt the chain of reasons and provide a way of solving Nagel's puzzle? The answer I think is that, when this happens, the puzzle is not yet solved. It seems a bit implausible to think that all, or even many, generous people are generous because they have adopted a policy of generosity for some reason. And of course generosity is only being used as one example here. The suggestion that character traits are often followed intentionally as matters of policy seems even less plausible if we think how many traits this would involve (being, e.g. trustworthy, loyal, helpful, friendly, courteous, kind, obedient, cheerful, thrifty, brave, fair, etc.), especially when we recall that it will also have to cover less attractive traits such as stinginess and vanity (as well as untrustworthiness, disloyalty, and do on).

So it seems plausible that in many, perhaps most, cases character traits will not be exemplified as matters of intentional policy.[15] In those cases the suggestion here will apply straightforwardly. And for at least some other cases, the reason the policy was adopted by the agent will itself be subject to the sort of suggestion about the role of character that we have been examining. If the fact that some action of mine would be a generous thing to do itself counted for me as a reason for doing it then, to that point at least, we still face the puzzle Nagel describes of why I performed the act for the reasons in favor rather than rejected it for the reasons against. So, the suggestion is, we can solve this puzzle (or at least, given the other issues surveyed above, take a first step toward solving it) by finding a character trait of mine that explains why I took

[15] Not to mention the fact that at least some character traits would seem not to allow an intentional policy at all. Modesty and shyness seem to be examples.

this fact (that it would be generous) as a reason for performing this action.

Are there other sorts of cases, cases where the suggestion we have been examining won't work because, so to speak, the agent's reasons "go all the way down" and there is no room, or perhaps need, for stepping outside the chain of reasons, in the way we have been discussing, to look at the character of the person who is doing the reasoning? Would this perhaps be the case of someone who looked with complete objectivity at all the relevant reasons, presumably many from the depths of moral philosophy, and then performed that action (or that kind of action) that objectively had the most reason on its side? For such a case to serve as an objection to the suggestion I am making the agent could *only* have character traits which were intentionally adopted; so that no appeal to any of her character traits would stop the "chain of reasons" in the way I have argued such appeals usually do. Is such a case possible? I am not sure. If it is though, then, in that case at least, the puzzle with which this paper began will remain unsolved.

Acknowledgements

An earlier version of this paper was presented as the Rod P. Dixon lecture at the University of Utah in November 2000. I am grateful to Michael Thompson, Michael Bratman, Elijah Millgram, Bruce Landesman, and Nick White for the questions, comments and suggestions they made on that occasion. Thanks are due as well to Sergio Tenenbaum for numerous helpful suggestions.

The argument of this paper has now appeared in an expanded and somewhat different version in Schueler (2003).

University of New Mexico
Philosophy Department
Albuquerque, NM 87131, USA
e-mail: Schueler@unm.edu

REFERENCES

Audi, R. (1989). *Practical Reasoning*. London: Routledge.

Davidson, D. (1980). How is Weakness of Will Possible? In: *Essays on Actions and Events*, pp. 21-42. Oxford: Oxford University Press.

Dennett, D. (1989). Three Kinds of Intentional Psychology. In: *The Intentional Stance*, pp. 43-82. Cambridge, MA: The MIT Press.

Hornsby, J. (1997). *Simple Mindedness*. Cambridge, MA: Harvard University Press.

Hume, D. (1975). An Enquiry Concerning the Principles of Morals. In: *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, pp. 169-323. Edited by P.H. Nidditch (3rd edition) and L.A. Selby-Bigge. Oxford: Clarendon Press.

Kim, J. (1993). The Non-Reductivist's Troubles with Mental Causation. In: J. Heil and A. Mele (eds.), *Mental Causation*, pp. 189-210. Oxford: Oxford University Press.

McGinn, C. (1979). Action and Its Explanation. In: N. Bolton (ed.), *Philosophical Problems in Psychology*, pp. 20-42. Methuen: London.

Nagel, T. (1986). *The View from Nowhere*. Oxford: University Press.

Ryle, G. (1949). *The Concept of Mind*. New York: Barns and Nobel.

Schueler, G.F. (2003). *Reasons and Purposes*. Oxford: Oxford University Press.

Williams, B. (1985). *Ethics and the Limits of Philosophy*. London: Fontana/Collins.

Michael Smith

IS THERE A NEXUS BETWEEN REASONS AND RATIONALITY?

When we say that a subject has attitudes that she is rationally required to have, does that entail that she has those attitudes for reasons? In other words, is there a deep nexus between being rational and responding to reasons? Many have argued that there is. Derek Parfit, for example, tells us that "to be rational is to respond to reasons" (Parfit 1997, p. 99). But I am not so sure. I begin by considering this question in the domain of theoretical rationality. The question in this domain is whether, when a subject has the beliefs that she is required to have by the norms of theoretical rationality, she is responding to reasons that there are for having those beliefs. Armed with a moderately clear answer to this question in the theoretical domain, I consider their relationship in the practical domain. When a subject has the desires that she is required to have by the norms of practical rationality, is she responding to reasons that there are for having those desires? Part of the interest of these questions lies in improving our understanding of reasons for action. I will say a little about this towards the end.

1. Reasons and Rationality in the Theoretical Domain

Let's begin with a very simple case of theoretically rational belief formation. Suppose that the following is true:

TR: Reason requires that if I believe that *p* and I believe that if *p* then *q*, then I believe that *q*.

Furthermore, suppose that I believe that *p*, and that I believe that if *p* then *q*, and that, abiding by TR, I rationally go on to form the belief that *q* – let's just assume whatever else needs to be true for this to be the case.

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 279-298. Amsterdam/New York, NY: Rodopi, 2007.

Should we conclude that, in such circumstances, there must be reasons for the formation of the belief that *q*? In other words, does the mere fact that I am theoretically rational in the formation of the belief that *q* entail that there are reasons to which I am responding?

It is important, in answering this question, that we do not move unwittingly between different senses of the term 'reason'. There is a widely accepted distinction in the philosophical literature between two senses in which we talk of reasons for action (Woods 1972, Smith 1987). On the one hand, some of our talk of reasons is talk about psychological states that are capable of rationally explaining an action. On the other, some of our talk of reasons is talk, not of psychological states that explain, but about considerations that justify. In these terms, the concern is that there is a similar distinction to be made in our talk of reasons for belief. Specifically, what we have asked appears to be ambiguous between a question about reasons for belief in the sense of psychological states that explain our beliefs, and reasons for belief in the sense of considerations that justify our beliefs. In answering the question we must therefore make explicit the sense of the term 'reason' that we have in mind.

Let's begin by considering reasons in the sense of psychological states that rationally explain. It is uncontroversial that, when I am theoretically rational in the formation of the belief that *q* in circumstances like those described, there are psychological states that explain my believing that *q*: I believe that *q because* I believe that *p* and I believe that if *p* then *q*, and it is in virtue of there being such a psychological explanation that I count as being theoretically rational. Since the nature of this psychological explanation will be important in what follows, let's say that in such circumstances the beliefs that *p* and that if *p* then *q evidentially explain* the belief that *q*. If the issue is whether there is a nexus between reasons in this sense and being theoretically rational then, in the circumstances described, the answer is that there is.

But what about reasons in the other sense? If, in circumstances like those described, I were asked what my reasons are for forming the belief that *q*, in the sense of the considerations that justify my belief, then I would unhesitatingly insist there are such reasons, namely, that *p* and that if *p* then *q*. But this doesn't yet entail that there are *in fact* reasons for my forming the belief that *q*, still less that the considerations I cite are reasons for doing so. Whether or not this is so turns on the relationship between the reasons that there are for forming the belief that *q* and the

reasons that I would cite as my reasons for forming that belief. Let's therefore consider that relationship.

When I cite the facts that *p* and that if *p* then *q* as my reasons for believing that *q*, it seems to me that I do so in full recognition of the possibility of there being a gap between the reasons that there are for believing that *q* and the reasons that I would cite. For one thing, I could readily admit that I have other reasons for believing that *q*, reasons that I don't, and perhaps even couldn't, presently know about. For another, I could readily admit that what I take to be reasons for believing that *q* are no such thing: that though I think that there are considerations that justify my believing that *q*, there aren't really.

As regards the first point, I can happily admit that there may be some way things are, say the *s* way, where if *s* then *q*, but where I am in no position to say whether or not *s*. In such circumstances it seems perfectly acceptable to say that there may be a reason to believe that *q* – namely, that *s* – which I am in no position to know about. Less dramatically, I can also quite happily admit that there may be evidence that is available to me, but to which I am not currently responsive, that supports *r*, where if *r* then *q*. In such circumstances it seems perfectly acceptable to say that my reasons for believing that *q* are more expansive than I currently think they are. Though I think that the only reasons I have for believing that *q* are that *p*, and that if *p* then *q*, I could thus accept the possibility that I have other reasons for forming the belief that *q* as well, namely, that *r*, and that if *r* then *q*.

Note that there is thus a crucial difference between two cases. In the former case, though we might happily admit that there are reasons for believing that *q* that I cannot know about, it seems to me that we would baulk at the suggestion that *I have those reasons* for believing that *q*, or that those reasons are *my reasons* for believing that *q*. When I say that certain considerations are *my reasons* for believing that *q*, or that *I have certain reasons* for believing that *q*, the use of these possessive expressions would thus seem to signal the availability of the reasons in question to me. I might just as well have said that there are reasons for believing that *q*, namely that *r* and if *r* then *q*, and that the facts that *r* and that if *r* then *q* happen to be available to me.

Consider now the second point: when I am asked what my reasons are for forming the belief that *q*, it seems that I could quite happily admit that what I take to be reasons for believing that *q* are no such thing. If, for example, I come to believe that it isn't the case that *p*, or that it isn't the case that if *p* then *q*, then I would say that though it seems to me that there are reasons for believing that *q*, there aren't really such reasons –

or, anyway, that no such reasons are known to me. This might of course be disputed. For how could the mere fact that my beliefs that *p*, or that if *p* then *q*, are false suffice to show that there is no reason for me to believe that *q*? But the very temptation to ask this question seems to me to depend on assuming what is at issue in this paper, namely, that when I am theoretically rational in believing what I believe, I believe what I believe for reasons.

To see why this assumption is false, we need to reflect a little on the relationship between the reasons that there are for forming beliefs and the beliefs that we form on the basis of those reasons. If my reasons for believing that *q* are that *p* and that if *p* then *q*, and if I believe that *q* for those reasons, then it would seem to follow that these reasons don't just justify my belief, but that they also figure in an explanation of my belief: I believe that *p because p* and if *p* then *q* (compare Williams 1980, p. 102). This is, if you like, what the 'for' signals in the claim that I believe that *q* for those reasons. It signals the presence of a relevant explanation. But, since all explanation is factive, it follows that, if these are indeed reasons for believing that *q*, then they must pick out ways that things really are. In other words, if my reasons for believing that *q* are that *p* and that if *p* then *q*, then it must be the case that *p* and that if *p* then *q* (Parfit 1997, p. 99; though contrast Dancy 2000, pp. 131-137).

This suggests that there is an important connection between the two sorts of reasons in the theoretical domain: that is, the reasons in the sense of the considerations that justify and the reasons in the sense of the psychological states that explain. Since, when there are considerations that justify a subject's forming the beliefs he forms, and the subject forms his beliefs on the basis of those considerations, the considerations themselves must figure in an explanation of his belief, and since, in such circumstances, there must also be psychological states that explain his belief, it follows that the two explanations of the subject's beliefs – that in terms of the considerations and that in terms of the psychological states – must somehow dovetail. And indeed they do. For what happens in such circumstances is this: the fact that *p* explains the subject's believing that *p*, and the fact that if *p* then *q* explains the subject's believing that if *p* then *q*, and this pair of beliefs explains the subject's believing that *q*. We will return to the connection between reasons in the sense of considerations that justify and reasons in the sense of psychological states that rationally explain, presently.

The two points just made suggest a preliminary story about the nature of reasons for belief, in the sense of considerations that justify. According to this story, the reasons that there are for believing that *q* –

the considerations that justify believing that q – are ways things are which make it rational for someone who believes that things are that way to believe that q. A subject's reasons for believing that q – or, equivalently, the reasons a subject has for believing that q – are a sub-set of these reasons. They are those reasons for believing that q that are available to that subject. The reasons for which a subject believes that q are a subset of this subset. They are those reasons for believing that q that are not just available to that subject, but which the subject is aware of and which figure in an explanation (of the right kind) of that subject's believing that q.

If some such story as this is along the right lines then it follows immediately that, if there is a nexus between reasons for belief, in the sense of the considerations that justify belief, and rationality at all, then that nexus is nowhere near as deep as we might at first have thought it was. There may be reasons for believing that q which no one knows about, and perhaps even that no one could know about. So though there are reasons for believing that q, it may be the case that no one is in a position rationally to form the belief that q. Conversely, even if various people are in a position rationally to form the belief that q, in so doing they might not be responding to reasons that there are to form that belief; they might not be believing that q for reasons. For though the rationality of their belief that q is guaranteed by the fact that their belief conforms to certain norms of theoretical rationality, the evidence available to them might be wildly misleading. There may be no reasons to believe that q, even though they are quite rational in forming that belief, and plainly do so for reasons in the other sense, that is, in the sense that there are psychological states that evidentially explain their belief.

With these observations in place we are in a position to answer the question with which we began. When a subject has the beliefs that she is required to have by the norms of theoretical rationality, it need not be the case that she is responding to reasons for having those beliefs in the sense of responding to considerations that justify. Such reasons for belief are, to repeat, ways things are which make it rational for someone who believes that things are that way to believe the thing in question. But since subjects can have evidentially well supported but wildly false beliefs, they are capable of rationally forming beliefs for no reason in this sense. What does seem to be true, of course, is that when a subject has the beliefs that she is required to have by the norms of theoretical rationality then she responds to what seem to her to be reasons to have those beliefs, in the sense of considerations that justify, and hence her beliefs are rationally explained – or better, evidentially explained – by

reasons, in the sense of psychological states. But, to repeat, this falls short of the requirement that there be any reasons for her beliefs, in the sense of there being ways things are that in fact justify her beliefs.

2. Is There a Nexus between Reasons and Rationality in the Practical Domain?

Let's now consider the relationship between reasons and rationality in the practical domain. We will focus on what many take to be a paradigm case of practical rationality, namely conformity to the means-ends norm (Smith 2004). Suppose that:

PR^{ME}: Reason requires that if I desire to and I believe that I can by -ing then I desire to .

Furthermore, suppose that I desire to , and that I believe that I can by -ing, and that, abiding by PR^{ME}, I rationally go on to form the desire to – again, let's just assume whatever else needs to be the case for this to be true. Should we conclude that, in such circumstances, since I conform to the norms of practical rationality, there must be reasons for the formation of the desire to ?

In answering this question it is once again important that we do not move in a slapdash fashion between two different senses of the term 'reason'. As before, the question we have asked is ambiguous between a question in the practical domain about reasons for desire in the sense of psychological states that explain our desires, and a question about reasons for desire in the sense of considerations that justify our desires.

Let's begin by focussing on the psychological states that explain desires. Once again, it seems that when a subject manifests the fact that she is practically rational by forming the desire to in the circumstances described, it does indeed follow that there must be reasons why she desires to in the sense of psychological states that figure in a rational explanation of her desire: she desires to *because* she desires to and believes that she can by -ing, and her being practically rational is crucially dependent on there being an explanation of her desire of just this kind. If the issue is whether there is a nexus between reasons in this sense and rationality in the circumstances described, then the answer is that there is. The situation is exactly the same as in the theoretical domain. It is, however, worth pausing to consider a difference in the nature of the rational explanations in these two domains.

Because the psychological states involved are quite different – beliefs, in the theoretical domain, pairs of desires for ends and beliefs about means to those ends, in the practical domain – it follows that the kinds of psychological explanation that we find in the two domains is quite different as well. In the theoretical domain we saw that the psychological states that rationally explain beliefs explain them in the sense of evidentially explaining them. But the psychological states that rationally explain desires, at least in the circumstances described, do not evidentially explain those desires (whatever that might mean – we will return to this presently). Instead they figure in what appears to be a straightforward teleological explanation of such desires.

To repeat, as we have just seen, there is indeed a nexus in the practical domain between reasons, in the sense of psychological states that rationally explain, and rationality. The question to ask next is whether there is a nexus between reasons and rationality in the other sense of 'reason'. When an agent is practically rational in circumstances like those described, must she be responding to reasons in the sense of considerations that justify?

What would I say if I was asked what my reasons are for forming the desire to in circumstances like those described? In other words, what considerations would I cite in giving a justification of my desire? Since one of my beliefs figures in the explanation of my desire to – since a belief of mine is one of the reasons, in the other sense of 'reason', for my desire – it follows that there is at least one consideration that I could cite, namely, the fact that I can by -ing. But we should not assume that I take this consideration to be any part of a set of considerations that justifies my desire to . For the truth is that I would be stuck if I was asked what the justification of my desire is. The question assumes that I believe there to be considerations that justify my formation of the desire to , whereas, at least as described, it seems that I need have no such belief. Worse still, on certain assumptions, there could be no such beliefs.

What would it be for there to be a consideration that justifies my desire to ? There would have to be a consideration that stands to my desire to in much the same relation as the considerations that justify my beliefs stand to the beliefs that they justify. As we saw above, such reasons for believing that *q* are the ways things are which make it rational for someone who believes that things are that way to believe that *q*. So, by analogy, reasons in this sense for desiring to – the considerations that justify desiring to – would have to be the ways things are which make it rational for someone who believes that things are that way to desire to .

But if this is right then the problematic nature of such reasons is readily apparent. If the norms of practical rationality are all like PR^{ME}, the norm requiring us to have desires for means when we have desires for ends and beliefs about the means to those ends, in the crucial respect of being requirements for us to have certain desires, given that we have certain other desires, then there is no way the world is such that someone's believing the world to be that way makes it rational for them to desire anything. Beliefs, all by themselves, are never enough to make it rational for people to have particular desires.

The conclusion is thus the extreme one that if all of the norms of practical rationality are like the means-ends norm in being requirements for us to have certain desires, given that we have certain other desires, then, when a subject has the desires that she is required to have by the norms of practical rationality, she is not responding to reasons for having those desires. There are no such reasons, and it is no part of our conception of what it is to be practically rational that agents take there to be such reasons either. Reasons in the sense of considerations that justify desires have nothing whatsoever to do with being practically rational.

3. Two Ways of Resisting the Extreme Conclusion in the Domain of Practical Rationality

In order to resist the extreme conclusion that reasons for desires, in the sense of considerations that justify having desires, have nothing to do with being practically rational, we would have to reject the assumption on which it is premised. The assumption is that the norms of practical rationality are all like the means-ends norm in being requirements for us to have certain desires, given that we have certain other desires. But on what grounds could we reject this assumption? There are two main strategies: the Besire Strategy and the Rationalist Strategy.

The Besire Strategy turns on the possibility of there being beliefs whose possession entails the possession of certain desires. Assume that PR^{ME} is true, and assume in addition that some claim of the following form is true:

BD: If an agent believes that *p* then she desires to .

With these assumptions in place we can derive the following principle:

PR^{BD}: Reason requires that if I believe that *p* and I believe that I can by -ing then I desire to .

PR^{BD} is a principle telling us which desires we are required to have given that we have certain beliefs. To be sure, the full story behind the truth of PR^{BD} goes via PR^{ME}, which is a principle telling us which desires to have given that we have certain other desires, and *BD*. But though this is an important fact about PR^{BD}, one which will be important to remember presently, it is plainly consistent with this that PR^{BD} is not itself – as it quite evidently is not! – a principle that tells us which desires to have, given that we have certain other desires. PR^{BD} tells us which desires to have on condition that we have certain beliefs.

But now suppose that PR^{BD} is true. It then follows that there could indeed be considerations which justify the formation of certain desires. For suppose that *p* and that I can by -ing. There is then a way the world is, namely, the way characterised by '*p*' and 'I can by -ing', which makes it rational for me to have a certain desire , namely the desire to , on condition that I believe that the world is indeed this way.

Nor is it hard to think of values for '*p*' and ' ' which, at least according to some theorists, makes a principle like PR^{BD} come out true. John McDowell has argued that when we believe that some person, *A*, is shy and sensitive – believe in the sense of fully understanding what it is for *A* to be shy and sensitive – then our having this belief entails our desiring to treat *A* in certain ways: to (say) protect *A* from those who would exploit her vulnerability (McDowell 1978). Let's call this:

BD^{McD}: If an agent believes that another person, *A*, is shy and sensitive then his having this belief entails his desiring to protect *A* from those who would exploit her vulnerability.

Suppose McDowell is right about BD^{McD}. Then, given PR^{ME}, we can derive the following norm of reason governing the formation of desires on the basis of beliefs:

$PR^{BD/McD}$: Reason requires that if I believe that *A* is shy and sensitive and I believe that I can protect her from someone who would exploit her vulnerability by -ing then I desire to .

So if what McDowell's says about the nature of the belief that someone is shy and sensitive is true, then there are considerations that justify the formation of certain desires to perform particular acts: that certain people are shy and sensitive and that a certain agent's acting in a particular way would protect them from those who would exploit their vulnerability are considerations that justify that agent's desiring to act in that particular way. These are reasons for the agent to desire to act in that way because

they are considerations which make it rational for him to desire to act in that way on condition that he believes that those considerations obtain.

It is important to note that the plausibility of the Besire Strategy as such isn't tied to the plausibility of McDowell's own view. The crucial question is not whether the belief that someone is shy and sensitive has the character McDowell claims it has, but rather whether there are any beliefs whose possession entails the possession of certain desires. If there are then, along the lines just sketched, we will be able to show that there could be reasons for the agent to have desires.

An alternative way to reject the assumption that the norms of practical rationality are all requirements for us to have certain desires, given that we have certain other desires, is to pursue the Rationalist Strategy. The Rationalist Strategy grants that there are no beliefs that entail the possession of particular desires. Instead it makes a normative claim. It tells us that there are additional basic norms of practical rationality, over and above PR^{ME}, norms of the following form that require us to have certain desires, given that we have certain beliefs.

PR^{R}: Reason requires that if I believe that *p*, then I desire to .

Unlike PR^{BD}, PR^{R} allows that an agent can have the belief that *p* without having the desire to . It simply tells us that that combination of belief and desire violates a norm of reason.

If some version of PR^{R} is true then it would once again follow that there could be considerations which justify the formation of certain desires. For suppose that *p*. There is then a way the world is, namely, the way characterised by '*p*', which makes it rational for me to have a certain desire , namely the desire to on condition that I believe that the world is indeed this way.

Nor is it hard to think of a value for '*p*' which, at least according to some theorists, makes a principle like PR^{R} come out true. According to one standard reading of Thomas Nagel's *The Possibility of Altruism* (1970), for example, Nagel argues that if someone believes that another creature is a person, equally real as himself, and he believes that that other person is in pain, and he believes that he can relieve that person's pain by acting in a particular way, then he violates a norm of practical reason if he doesn't desire to act in that particular way (Wallace 1990; for an alternative reading of Nagel according to which he is pursuing the Besire Strategy, see Smith 1994, Ch. 4). The following is thus, by Nagel's lights, a self-standing norm of practical rationality alongside the means-ends norm:

$PR^{R/N}$: Reason requires that if I believe that another creature, *A*, is a person, equally real as myself, and I believe that *A* is in pain, and I believe that I can relieve *A*'s pain by -ing, then I desire to .

So, if what Nagel says is true, then there are considerations that justify the formation of certain desires to perform particular acts: that another creature is a person, equally real as an agent himself, and that that person is in pain, and that the agent himself can relieve that person's pain by acting in a particular way, are considerations that justify the agent's desiring to act in that particular way. These are reasons for the agent to desire to act in that way because they are considerations which make it rational for him to desire to act in that way on condition that he believes that those considerations obtain.

As with the Besire Strategy, it is important to note that the plausibility of the Rationalist Strategy as such isn't tied to the plausibility of Nagel's view. The crucial question is not whether the beliefs Nagel cites have the characteristics he says they do, but rather whether there are any beliefs whose possession makes it rational for an agent to have certain desires. If there are then, along the lines just sketched, we will be able to show that this entails the possibility of reasons for the agent to have desires.

Though the two strategies both underwrite the possibility of there being reasons for desires, in the sense of considerations that justify having desires, as we have already noted in passing, there is a significant difference in the way in which the two strategies underwrite this possibility. This difference underscores a difference in the two strategies' views about the nature of the psychological states that explain desires, and hence a difference in what the two strategies tell us about the nature of reasons for desires in the other sense of 'reason', the sense in which reasons are psychological states that explain desires.

To repeat, the Besire Strategy tells us that the desires that are justified are produced in the normal means-end way: $PR^{B/McD}$ is derived from BD^{McD} and PR^{ME}. The agent has a belief that entails that he has a desire for an end, and this desire for the end combines with a belief about a means to that end to produce a desire for the means in the normal means-end way. The desire that is justified – the desire for the means – is thus susceptible to a regular teleological explanation. The reason for the desire, in the sense of the psychological state that explains that desire, is thus a psychological state that is suited to play the required role in this teleological explanation. It is a desire, albeit a desire whose possession happens to be entailed by the possession of a belief.

However the Rationalist Strategy, by contrast, tells us that desires that are justified are produced in a completely different way: $PR^{R/N}$ is supposed to be a basic norm of practical rationality alongside PR^{ME}. The agent has a belief which produces the relevant desire, but this belief does not produce that desire by combining with a desire for an end. The desire that is justified is thus not susceptible to a regular teleological explanation, and the reasons for the desire, in the sense of the psychological states that explain the desire, that are posited by the Rationalist Strategy are thus not psychological states that are well suited to providing a teleological explanation. But in that case, what sort of explanation is it?

It is irresistable, I think, to conclude that the reasons for desires, in the sense of the psychological states that explain desires, posited by the Rationalist Strategy are supposed to explain desires in a way that is strongly analogous to the way in which the reasons, in the psychological state sense, for belief explain beliefs. In other words, notwithstanding the fact that it is an explanation in the practical domain, the style of explanation is strongly analogous to evidential explanation. Of course, the explanation is not evidential explanation pure and simple, for the beliefs in question do not support the truth of the desire: desires aren't the sort of state that can be true or false. But the explanation is strongly analogous to evidential explanation in so far as the beliefs in question suffice to make the desire in question the one that it is sensible to have independently of what other desires are had. In this extended sense the beliefs bear evidentially on the desires: the beliefs mandate possession of the desires all by themselves; to desire otherwise is to fly in the face of the facts.

4. Is Either of the Ways of Resisting the Extreme Conclusion Plausible?

Is either of these strategies plausible, and, if so, which is more plausible? Though this is not the place to argue the point at great length, my own view is that the extreme conclusion is unstable. Even if we begin just by assuming that norms of practical reason tell us which desires to have, given that we have certain desires, we are quickly led from this to the conclusion that there are norms telling us which desires to have, given that we have certain beliefs. Moreover, we are led to this conclusion in the manner suggested by the Rationalist Strategy, not the Besire Strategy, for the Besire Strategy is hopeless.

The problem with the Besire Strategy, in a nutshell, is that it requires a far too demanding account of what it takes to understand a proposition. In many ways, these flaws are analogous to flaws in the view of belief according to which belief is closed under *a priori* consequence. Since I take it that few people hold that view about belief, let me give a brief reminder of its flaws first, and then I will spell out the analogy. It will then be clear not just why we should reject the Besire Strategy, but why we should focus our attention on the Rationalist Strategy.

Suppose *p a priori* entails *q* and consider the principle:

BB: If a subject believes that *p* then she believes that *q*.

This principle is immensely implausible, on the face of it, for the fairly flat-footed reason that it is one thing to understand and believe that *p* and quite another to see that *q* is an *a priori* consequences of *p*. This is not, of course, to deny that there may be some *q*s for which BB is plausible. But the *q*s for which it is plausible are precisely those for which it can be established that what is not in general true is true at least in this particular case: someone who both believes and understands that *p* believes this particular *a priori* consequence of *p*, namely, *q*. An example will help bring out this flat-footed point.

Mathematicians debate long and hard about whether various mathematical conjectures are true or false. But, if true, these conjectures are, let's suppose, *a priori* consequences of various other more basic mathematical propositions that the mathematicians claim to believe. Now suppose that a particular mathematical conjecture is true. Should we conclude that the mathematicians who claim to believe these more basic mathematical propositions are either speaking truly, in which case they already believe the conjecture, notwithstanding the fact that they claim that they don't know whether the conjecture is true or false, or that they are speaking falsely because, since they are right that they don't know whether the conjecture is true or false, it follows that they mustn't really understand the more basic mathematical propositions from which the conjecture follows *a priori*? If, like me, you find it extremely hard to believe either of these things then you have no choice but to agree that BB is implausible.

But what is the alternative to BB? The alternative is to suppose that, when *q* is an *a priori* consequence of *p*, and when the *q* in question is not one of *q*s just mentioned, the belief that *p* and the belief that *q* stand in the following normative relation:

TR^{APC}: Reason requires that if a subject believes that *p* then she believes that *q*.

TR^{APC} is a principle of theoretical rationality which tells us that though it is possible for someone to believe that *p* without believing that *q* when *p* is an *a priori* consequence of *q*, someone who has this pair of beliefs violates a norm of reason. Mathematicians who believe the more basic mathematical propositions from which a true conjecture follows *a priori* violate a norm of reason in failing to believe the conjecture. This is a very plausible claim. TR^{APC} is thus far more plausible than BB.

The analogy between the view that belief is closed under *a priori* consequence and the view of the relationship between belief and desire propounded by those who pursue the Besire Strategy is, I hope, already apparent, but in case it isn't, let me spell it out. The following two principles are extremely similar in crucial respects:

BB: If a subject believes that *p* then she believes that *q*.
BD: If an agent believes that *p* then she desires to .

For both BB and BD require us to have extremely high standards of what it takes to understand a proposition. And the following two principles are also extremely similar in crucial respects:

TR^{APC}: Reason requires that if a subject believes that *p* then she believes that *q*.
PR^{R}: Reason requires that if an agent believes that *p* then she desires to .

For just as reflection on the inadequacy of BB forces us to focus on the more plausible TR^{APC}, so reflection on the inadequacy of BD forces us to focus on the more plausible PR^{R}. Though there do not seem to be any values of '*p*' and ' ' that make BD come out true, it would seem to be at the very least an open question whether there are any values of '*p*' and ' ' that make PR^{R} come out true.

In fact, it seems to me that, on the basis of fairly uncontroversial premises, we can provide an argument that closes the question in favour of there being at least one instance of PR^{R}. Suppose, *pro tem*, that:

PR^{ME}: Reason requires that if I desire to and I believe that I can by -ing then I desire to ,

is the only norm of practical reason. Now suppose that an agent believes that, if she had a desire set that conforms perfectly to all of the norms of reason, then she would desire that she s, but that she does not desire to . It seems to me that we can now provide an argument for the claim that the following is a norm of reason:

PR^{R*}: Reason requires that if an agent believes that she would want that she s if she had a desire set that conforms to all of the norms of reason, then she desires that she s,

and, hence, that PR^{ME} is not the only norm of reason.

Moreover, PR^{R*} is not just an additional norm of reason, it is a norm of reason of the same form as PR^{R}. In other words, it is a norm which requires agents to have certain desires, given that they have certain beliefs. So, starting with just the assumption that PR^{ME} is a norm of reason, we see that, given agents can form beliefs about what they would want if their desires conformed to norms of reason, there are indeed norms of reason of the same form as PR^{R}. Let me say a little in support of these claims.

For PR^{R*} to be plausible, it is important that we focus on the right logical form of the belief in question. The idea is not that the agent believes that, in the nearest possible world in which her desire set conforms perfectly to the norms of reason, she desires that she s in *that* world. Reason certainly doesn't ban an agent's failing to desire to do *in this world*, in which she does not conform to all of the norms of reason, something that she believes that she would desire herself to do *in that world*, in which she does. For the believed difference in her circumstances – the fact that she believes that she conforms to the norms of reason in that world, but not in this world – might well make all the difference to what it would be sensible for her to want to obtain in that world as opposed to this. Rather, the idea is that the agent in question believes that, in the nearest possible world in which her desire set conforms perfectly to the norms of reason, she desires that she s *in this world*.

Consider an example. Suppose I believe that I desire the creature comforts, and that I can have the creature comforts by earning money, but I also believe that, because I am means-end irrational – remember, PR^{ME} is the only norm of practical reason that we are admitting at this stage – I don't desire to earn money. Moreover suppose that my belief that I don't desire to earn money is true. In that case, when I consider what I would want myself to do in this world in the nearest possible world in which my desire set conforms perfectly to the norms of reason, I conclude that what I would want myself to do in this world is earn money. I come to this conclusion because, in that world – the world in which my desires conform to all of the norms of practical reason – my desires conform to PR^{ME}. So the upshot is that I believe that I would desire myself to earn money in this world if I had a desire set that

conformed perfectly to the norms of reason, but that I don't desire to earn money in this world.

To repeat, it seems to me that reason is against this pairing of belief and desire. The explanation of this ban is relatively simple. I am, after all, a creature who is not just capable of having beliefs and desires that are subject to norms of reason, but a creature who is also capable of forming views about which beliefs and desires I would have if my beliefs and desires conformed to norms of reason, and a creature who is capable, as well, of managing my beliefs and desires in the light of the views I form. In these terms, the problem with my lacking a desire to earn money in the circumstances described is that it shows the extent to which I fail in that management role. It shows that I fail in that management role in an exactly analogous way to the way in which I would fail in that management role if I failed to believe that *q* when I believe that I would believe that *q* if I had a belief set that conforms to all of the norms of theoretical reason (for an alternative view see Sayre-McCord 1997).

Note that the argument just given doesn't assume anything controversial about the nature of norms of reason. In particular, it doesn't presuppose that there are any other norms of reason of the same form as PR^{R} beyond PR^{R*}: that is the reason the example was given while assuming, *pro tem*, that PR^{ME} is the only norm of practical reason. For all that the argument just given tells us, intrinsic desires might therefore one and all be rationally optional. That is neither here nor there, as regards the argument just given. For what that argument seems to establish is that, beginning from a very minimal assumption about the nature of the norms of practical reason, an assumption congenial to the extreme view – the assumption that PR^{ME} is a norm of practical reason that governs desires – once we fully internalise the consequences of the fact that we can form beliefs about what we would want if our desires conformed to this norm, and the fact that we can manage our desires in the light of these beliefs, we see that we there is a more substantive norm governing our desires, PR^{R*}, which is a norm of the form PR^{R}. The upshot is thus that we must reject extreme view described earlier. We have no alternative but to pursue the Rationalist Strategy.

5. What Have We Learned about Reasons for Action?

I said at the outset that part of the aim of this paper is to improve our understanding of reasons for action. In this final section I will say a little about what we have learned from the previous discussion.

As I have already said, it is widely accepted that we can distinguish two senses in which we talk of reasons for action. Some of our talk of reasons for action is talk about the psychological states that are capable of rationally explaining actions, and some is talk, not of psychological states that explain, but of considerations that justify actions. Let's call reasons of the first sort "motivating" reasons, and reasons of the second sort "normative" reasons (Smith 1994, Ch. 4).

As is no doubt obvious, the argument just given for PR^{R*} is, in effect, an argument in favour of the conception of normative reasons proposed by Bernard Williams in his famous paper "Internal and External Reasons" (Williams 1980; see Pettit and Smith forthcoming). One consequence of our discussion, then, is that we must suppose that there are at least some normative reasons. Though the extreme view that there are no considerations that justify desires looks like it might well have implied that there are no normative reasons for action either, that extreme view is unstable. There is at least one consideration that justifies desire, namely, the fact that the desirer herself would want herself to act in the way desired if she had a desire set that conforms to all of the norms of practical reason. Whether there are further considerations that justify desires is an open question, one which I hope to pursue elsewhere.

As regards motivating reasons, the standard theory is the one we have inherited from Hume, a theory which has more recently been popularised by Donald Davidson (Hume 1740; Davidson 1963). According to this theory, an agent's motivating reasons rationally explain his action in a characteristic teleological manner: motivating reasons embody the goals that the agent has in acting and they embody his conception of what he is doing as something that will achieve his goals. Hume tells us that motivating reasons are psychological pairs comprising an agent's desires for ends (these embody his goals) and means-end beliefs (these embody his conception of what he is doing as something that will achieve his goals), where desire and belief, in turn, are distinct existences: no matter what combination of belief and desire an agent has, we can always imagine an agent who has those beliefs and yet who has different desires, and vice versa. Moreover, the fact that we can always find motivating reasons – that is, desire and belief pairs – that teleologically explain an action is taken by the standard theory to be constitutive of action: what makes an event an action is the fact that it is something that is done for a reason, in the sense that there is a motivating reason, a desire and means-end belief pair, that teleologically explains it.

It perhaps goes without saying that the extreme conclusion mooted above sits happily alongside the Humean theory of motivating reasons.

The extreme conclusion, you will recall, was that though there may be reasons for desires in the sense of psychological states that rationally explain them – desires can, after all, be teleologically explained by other desires – there can be no reasons for desires in the sense of considerations that justify them. This sits happily alongside the Humean theory of motivating reasons because it simply adds a detail to that theory. It tells us that though there may be considerations that justify the means-end beliefs that partially constitute motivating reasons, there are no considerations that justify the desires for ends that partially constitute motivating reasons. But what if we were to reject that extreme conclusion? Suppose we could successfully pursue either the Besire Strategy or the Rationalist Strategy. Would this force us to adopt a different view about the nature of motivating reasons?

The Besire Strategy challenges a core element in the standard Humean theory of motivating reasons, for it takes issue with the idea that belief and desire are distinct existences. Moreover, because it takes issue with this core idea it must reject the account of what makes an action an action. Those actions that are teleologically explained by desires whose possession is entailed by beliefs the agent has are plainly not actions in virtue of being susceptible to a teleological explanation by desires that are distinct from beliefs. So if we can successfully pursue the Besire Strategy, then we must reject the standard Humean theory of motivating reasons. But since, as I have already explained, the Besire Strategy is hopeless, this challenge doesn't seem to me to be of much concern.

The Rationalist Strategy, by contrast, provides no real challenge to the standard Humean theory of motivating reasons. For it takes the idea that belief and desire are distinct existences for granted and simply adds a detail. According to the Rationalist Strategy, the desires that partially constitute motivating reasons can themselves be rationally explained by beliefs, rationally explained in a sense strongly analogous to evidential explanation. Though Hume himself didn't believe that desires were susceptible to such rational explanation, it is plainly consistent with the Humean theory of motivating reasons, as outlined above, that Hume was wrong about this and desires are suceptible to such rational explanation. Moreover it is also plainly consistent with the possibility of giving such a rational explanation of desire that the fact that this is so is in no way essential to an action's being an action. What makes an action an action is the fact that it can be teleologically explained by a desire for an end and a belief about means, something we can establish to be so under a conspiracy of silence about the rational etiology of that desire and belief.

Whether or not we can successfully pursue the Rationalist Strategy is thus orthogonal to the Humean theory of motivating reasons.

6. Conclusion

We are now in a position to answer the quite general question that we asked at the beginning of this paper. When we say that a subject has attitudes that she is rationally required to have, does that entail that she has those attitudes for reasons? In other words, is there a deep nexus between being rational and responding to reasons?

The answer to this question is a quite decisive "No." A subject's having the beliefs that she is required to have by the norms of theoretical rationality entails, at most, that she responds to the reasons that it seems to her there are for forming that belief, not the existence of such reasons. And, on a certain crucial assumption, a subject's having the desires that she is required to have by the norms of practical rationality entails neither the existence of reasons for forming that desire nor that it seems to her that there are such reasons for forming that desire. Unlike theoretical rationality, practical rationality is not even a matter of being responsive to what seem to be reasons.

However, to repeat, this is so only on a certain crucial assumption, namely, that the norms of practical rationality are all like the means-ends norm in being requirements to have certain desires, given that we have certain other desires. There are two main strategies that can be pursued in the attempt to reject this assumption: the Besire Strategy and the Rationalist Strategy. The Besire Strategy requires us to find values of '*p*' and ' ' that make a principle of the following form come out true:

BD: If an agent believes that *p* then she desires to .

The Rationalist Strategy requires us to find values of '*p*' and ' ' that make a principle of the following form come out true:

PR^R: Reason requires that if an agent believes that *p*, then she desires to .

Each strategy, if successfully pursued, would provide us with an account of how there could be reasons for desires in the sense of considerations that justify our having certain desires.

However, I have also explained why, as it seems to me, the Besire Strategy is not very plausible, and why we are bound to have at least moderate success when we pursue the Rationalist Strategy. There is at

least one consideration that justifies desire, namely, the fact that the desirer herself would want herself to act in the way desired if she had a desire set that conforms to all of the norms of practical reason. Whether there are further considerations that justify desires is an open question, one which must be pursued on another occasion.

Princeton University
Department of Philosophy
1879 Hall
Princeton NJ 08540, USA
e-mail: msmith@princeton.edu

REFERENCES

Dancy, J. (2000). *Practical Reality* Oxford: Oxford University Press.

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy* **60**, 685-700. Reprinted in: *Essays on Actions and Events*, pp. 3-20. Oxford: Oxford University Press, 1980.

Hume, D. (1740). *A Treatise of Human Nature.* Oxford: Clarendon Press.

McDowell, J. (1978). Are Moral Requirements Hypothetical Imperatives? *The Aristotelian Society Supplementary Volume* **52**, 13-29.

Nagel, T. (1970). *The Possibility of Altruism.* Princeton, NJ: Princeton University Press.

Parfit, D. (1997). Reasons and Motivation. *The Aristotelian Society Supplementary Volume* **71**, 99-130.

Pettit, P. and M. Smith (2006). External Reasons. In: C. Macdonald and G. Macdonald (eds.), *McDowell and His Critics*, pp. 142-169. Oxford: Blackwell.

Sayre-McCord, G. (1997). The Meta-Ethical Problem: A Discussion of Michael Smith's *The Moral Problem*. *Ethics* **108**, 55-83.

Smith, M. (1987). The Humean Theory of Motivation. *Mind* **96**, 36-61.

Smith, M. (1994). *The Moral Problem.* Oxford: Blackwell Publishers.

Smith, M. (1997). In Defence of *The Moral Problem*: A Reply to Brink, Copp and Sayre-McCord. *Ethics* **108**, 84-119.

Smith, M. (2004). Instrumental Desires, Instrumental Rationality. *The Aristotelian Society Supplementary Volume* **78**, 93-109.

Wallace, J.R. (1990). How to Argue about Practical Reason. *Mind* **99**, 267-297

Williams, B. (1980). Internal and External Reasons. In: *Moral Luck: Philosophical Papers 1973-1980*, pp. 101-114. Cambridge: Cambridge University Press.

Woods, M. (1972). Reasons for Action and Desire. *The Aristotelian Society Supplementary Volume* **46**, 189-201.

David Sobel

PRACTICAL REASONS AND MISTAKES OF PRACTICAL RATIONALITY

This paper will consider a broad objection against subjective accounts of reasons for action. I will conclude, tentatively, that there is no telling objection in the neighborhood where many have thought there was. The objection is this. Subjectivists have generally been clear that the concerns that allegedly determine one's practical reasons are counterfactual concerns. Typically, for example, it is held that such concerns must be informed. But subjectivists have had little to say about what else must go right besides having good information if the resulting concerns are to have normative authority. Yet it seems possible to make bad uses of good information. If such practical processing of good information can be done well or badly, then subjectivists would have to tell a story about what counts as good practical processing of good information. The worry is that subjectivists cannot tell an adequate story about the differences between good and bad practical processing. And if this were so, then subjectivists would be unable to provide a convincing story about our reasons for action. As I am understanding this worry, it purports to be a preemptive criticism of all versions of subjectivism such that, if it were successful, one could be assured that no subjectivist account could be acceptable.

As I say, I will attempt to defend subjectivism from this worry. But to be honest, I am still struggling with how best to understand this underdiscussed and important general criticism of subjectivism. Further, a full response to such a worry would probably require a fully worked out version of subjectivism and I do not have such a theory to offer and I do not think anyone else does either. Here I aim to make a start at defusing the preemptive worry.

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 299-321. Amsterdam/New York, NY: Rodopi, 2007.

1. Subjectivism and Concerns

Subjective accounts of reasons for action claim that what makes it the case that consideration *C* provides *P* a reason to *O* is the existence of certain contingent features of agency, namely concerns, of *P*. It is the contours of a person's contingent concerns that makes it the case that she does or does not have a reason to do something.[1] The subjectivist claims that ideal deliberation, absent input from an agent's contingent concerns, does not yield determinate conclusions about what the agent has reason to do. Thus, for example, the fact that I happen to like Dairy Queen swirl cones dipped in chocolate might give me a reason to stop there on my way home from work on such an account. But if I did not happen to like that, I would not have such a reason.

The notion of concerns at work here is important to the subjectivist account but difficult to satisfactorily analyze. I will not have enough to say about such states here. The most popular neo-Humean understanding of desires is based on the notion of direction of fit. Intuitively the idea is that beliefs aim to track the world (and so tend to go out of existence when one is confronted with an appearance to the effect that the world is not as one believes) whereas desires aim not to track the world but to impose themselves upon the world (such that they do not similarly tend to go out of existence in the face of such appearances). I think the direction of fit understanding of desires is unsuccessful, yet I lack a better account.[2] Essential to the subjectivist's understanding of concerns (wants, preferences, desires, etc.) is that such states are not truth-assessable and not best thought of as accurate or inaccurate responses to the value of options.[3] To take a simple example, I can prefer chocolate

---

[1] This contrasts with the general understanding of what Stephen Darwall has called, variously, "existence internalism" or "metaphysical internalism" which is presented, by Darwall, Williams, Korsgaard and others, as a thesis about necessary or necessary and sufficient conditions for being a reason. So understood, the thesis of internalism does not commit one to a view about what makes it the case that *C* provides *P* with a reason to *O*. Thus the internalism debate addresses a different question than the subjectivism debate and one could be an internalist objectivist. Indeed, I think that Michael Smith's view of reasons in (1994) counts as an instance of internalist objectivism. For an argument that we ought to reject internalism but can nonetheless embrace subjectivism, see my (2001a). For related arguments see Johnson (1999).

[2] Michael Smith, in his (1994), offers a good presentation of this view. For criticism of such views see Sobel and Copp (2001, pp. 44-53).

[3] On an increasingly popular conception of desire, to have a desire for *P* constitutively involves taking oneself to have a reason for *P*. Subjectivists must reject such a view. For a prominent recent advocate of the subjectivism unfriendly understanding of desire see

ice cream to vanilla or have the reverse preference, without making any mistake about either flavor or their respective value.

In the case of belief it is typically thought that the truth provides standards that belief should aim for (some say that only if a state aims at truth is it a belief). The subjectivist denies that there is any comparable standard for desire such that desires would be correct or accurate if they matched up with that standard. As the subjectivist sees it, the value of the various options for an agent is determined by that agent's informed concerns; not already there to guide such concerns. Putting it this way combines the two central subjectivist theses. The first is that there are no value facts that should (or could) serve as standards for desire. The second is that, suitably dressed up, desires can have (or be the only source of) practical normative authority.

2. Subjectivists' Focus on Errors of Theoretical Reason

However we understand them, an ordinary person will have, at best, imperfect access to the concerns that subjectivists most plausibly take to be relevant to her reasons. The satisfaction of some of our actual introspectable concerns can leave the taste of dust in our mouths. Thus it has become standard for subjectivists to offer a counterfactual analysis of the vantage point from which it is alleged that our concerns determine our reasons.[4] Our subjectivist must suggest that there is a way we could be such that our attitudes determine our reasons.[5] And once on this path there is a tendency for the analysis to become wildly counterfactual to the extent that a typical view now would suggest that our relevant attitudes at least have to be shaped in the light of complete factual information about the universe.[6]

Scanlon (1998, Ch. 1). For arguments against this view see Copp and Sobel (2002, especially pp. 269-272). See also Scanlon's reply to this paper in (2002).

[4] Interestingly, while different subjectivists argue for different vantage points, they each tend to offer a single description of this vantage point such that for each agent she only counts as properly situated if she gets herself into that vantage point. There is room to wonder if this one-size fits all understanding of the vantage point is compatible with the commitments of subjectivism. See Connie Rosati (1996, pp. 247-273) for an interesting discussion and an alternative to the one-size fits all model.

[5] Donald Hubin (1996, pp. 31-54) offers a subjectivist position that claims that our current actual intrinsic concerns determine our reasons. I take it, however, that Hubin allows that we have only imperfect access to such concerns.

[6] This tendency might be thought to start with Mill's competent judges test (which is admittedly offered as an account of well-being, not reasons) and run through Sidgwick (1981, pp. 111-122), Brandt (1979, pp. 10, 113, 329), Hare (1981, pp. 101-105 and

The wilder the counterfactuals become, the clearer it is that such accounts do not aspire, in the first instance, to tell people what kind of thoughts should enter into their heads in everyday practical reasoning. Just as consequentialism is best understood as a theory about the truth-maker of moral claims, and thus is compatible with recommending a decision-procedure for ethical situations that does not simply mimic the thought process invoked at the level of truth-maker, so too can this happen in the case of subjectivist accounts of reasons for action.

Understanding subjectivism this way we can say, seemingly with Hume and Williams, that an agent who deliberates sensibly given her epistemic situation might nonetheless fail to act according to her genuine reasons because her deliberation involved factual errors. Such rational but not ideally informed deliberation might fail to lead one to see one's genuine reasons even if one's factual premises were mistaken in a way that is not culpable.[7] On such a conception, it is clear why pride of place would be given to factual mistakes in explaining how an agent might come to act contrary to her true reasons.

And indeed, influential subjectivist accounts of reasons for action, such as those nearly offered by Hume and Bernard Williams, when they explain how an agent could act contrary to her true reasons for action, tend to focus attention on cases in which the agent has been misled by false factual information. Hume discusses an example in which a fruit-fancier acts contrary to reason in trying to get a certain fruit because she does not realize that the fruit is rotten (Hume 1967, p. 460). Williams fixes on a case in which a person acts contrary to her reasons in drinking a petrol and tonic because the agent falsely believes that what she is about to drink is a gin and tonic (Williams 1981, pp. 102-103). In neither of these cases do these authors provide any reason to suspect that the mislead agents were foolish or gullible in deciding to take these actions. The main problem in these cases appears to be merely that the agents had, perhaps non-culpably, false beliefs.

214-216). See also Senor and Fotion (1990, pp. 217-218), Williams (1981) as well as his (1995), Griffin (1986, pp. 11-17), Rawls (1971, pp. 407-424), Gauthier (1986, Ch. 2), Darwall (1983, Part II), Harsanyi (1982, p. 55), Railton (1986, pp. 5-31), Lewis (1989, pp. 113-137), Kagan (1989, Ch. 8). Several important caveats apply to some of the above author's commitments to subjectivism and some would decline the label.

[7] I argued that we should so understand subjectivism, or at least the strand of subjectivism about reasons for action that people connect with Hume and Williams in (2001). See also Railton (1997, especially pp. 60-61 and 77-78) and Hubin (1996, p. 51).

3. The Worry

Generally, subjectivists have focused on the information component of ideal deliberation, not paying adequate attention to other aspects of ideal deliberation.[8] For example, Peter Railton's quasi-official statement of his account of a person's good runs like this: "an individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality." I take it that the "lapses of instrumental rationality" clause only has teeth after the account has determined what the agent wants in the relevant way. That is, on such an account it is the special sort of concerns that set the end, and only after the end is set is it possible to fail to be instrumentally rational. Further, I take it that the "cognitive error" clause speaks to the agent's informational input rather than to processing issues (Railton 1986, p. 16).[9]

Hume at times denied the need for an account of good processing, suggesting that once the information component was ideal the processing would take care of itself (Hume 1967, p. 416). But Williams suggests that the subjectivist should have something to say not only about the truth of the ideal deliberator's beliefs, but also about how she processes her true beliefs in generating her concerns. For example, Williams claims that the sound deliberator must use her imagination, and presumably he means use her imagination well, which is something she might fail to do. So Williams presumably sees the need for an account of good information processing which will include at least an account of good uses of imagination.[10]

---

[8] Some have claimed that subjectivists lack resources to motivate even the information requirement. I will here ignore this complaint.

[9] I do not mean to suggest that Railton utterly ignored the possibility of processing errors. He does, however, allow that his focus is "on the problem of full information rather than full rationality" (Railton 1986).

[10] Williams offers a laundry list of examples of good processing that the sound deliberator should go in for in (1981, pp. 104-105). I take it that each example has a flip side. That is, each offers an example of ways that the deliberator could have processed poorly but did not. Thus each of Williams's examples serves as a counter-example to Korsgaard's claim that the Humean cannot make sense of less than ideal reasoning. Korsgaard's claim that the subjectivist cannot offer a normative account of rationality or ideal rationality ignores the fact that Williams offers exactly that. She seems not to keep an eye out for the ways in which Humeans can and do differ from Hume. Of course if each of Williams's examples were susceptible of being understood as really just an error of theoretical reason this would undermine the thought that Williams's examples refute Korsgaard's claim. My

But there are several concerns one might have with Williams's account of good processing. First, he offers only headings under which he claims there can be good or bad processing, he does not himself show us how to distinguish between good and bad processing under these headings. He does imply that the subjectivist can make sense of and underwrite the distinction between good and bad kinds of processing, but this is merely claimed, not demonstrated. Second, and somewhat related to the first worry, Williams's account of reasons is remarkably vague. Williams's efforts are only to offer an account of pro tanto reasons, not to give any account of the relative strength of reasons. Further, he officially offers only a necessary condition for having pro tanto status. Thirdly, it is not clear that Williams's account of the headings under which the subjectivist should say that there can be good and bad processing are correct. It seems that many, if not all, of Williams's headings point us to mistakes that are best understood as mistakes of theoretical reason rather than mistakes of processing.

Thus, for example, it might be claimed that Williams' favored case of excellence in processing, namely good use of imagination, is just standing in for a full appreciation of the facts, propositional and phenomenological, together with a lively appreciation of the full variety of options that are truly available at various junctures. And a failure here is just a failure of ideal theoretical reason. Thus, although Williams seems to recognize the need for an account of ideal processing, he offers little help in developing such an account. Still less should we be persuaded that Williams has identified and offered fixes for the full range of possible errors in processing.

In sum, the best subjectivist-friendly accounts of ideal deliberation have been much clearer and helpful in developing the information component of ideal deliberation and have largely ignored the challenges involved in developing an account of ideal processing.

But it is surely intuitively quite plausible to think that there must be another way for an agent to act contrary to her reasons besides not knowing the facts. It seems that we must be able to make sense of the possibility of an agent knowing the facts yet still making mistakes in her deliberation. Cases of weakness of will are merely the most obvious example. It seems we need a second way that a person can act contrary to her reasons besides merely having poor informational input into deliberation. An adequate account of practical reason, it would certainly

---

point is merely that it is odd that Korsgaard did not feel called upon to make out a case that Williams list of examples of faulty processing does not put pressure on her claim that the subjectivist can offer no such examples.

seem, must be able to make sense of the possibility of poor processing of good information. Only when good information is combined with good processing should we have any confidence that the resulting concerns determine the agent's reason. If there can be such a thing as faulty processing of good information, then a subjectivist theory would have to ensure that none of this went on in generating the concerns that are alleged to determine an agent's reasons.

Thus the availability of the general claim that an adequate account of practical reason must make room for the possibility of practical errors in processing, but that subjectivist accounts are uniquely ill situated to provide for such a possibility. This paper will principally be an investigation into the force of this critique of subjectivism.

One reason a subjectivist seems ill positioned to distinguish between good and bad processing is that the subjectivist is committed to the idea that no end is so crazy that a person could not have a reason to go in for such a thing. Thus the subjectivist must not say that any processing that leads to the conclusion that one should spend one's life counting blades of grass is necessarily faulty. Subjectivists cannot tar action as contrary to reason simply because it is aimed at any particular target. Another explanation for the supposition that the subjectivist cannot offer an adequate account of processing errors is that the subjectivist is sometimes held to be committed to the thought that the only way information can appropriately impact on one's desires is causally. Thus the subjectivist could not say that a bit of true information had such and such a causal impact on an agent's desires, but that it ought not to have had such an effect.

On the other hand, it is unclear and under-discussed what forms of bad processing exist once we ignore or overcome all errors of theoretical reason. The complaint against subjectivism that it cannot capture the distinction between good and bad processing requires both explicating a non-question-begging (or at least obvious) example of poor processing and arguing that subjectivists lack the resources to accommodate adequately the possibility of such poor processing.[11]

---

[11] The issue of what resources are available to the subjectivist is vexing because it requires understanding what moves are continuous with (or at least not at odds with) the general subjectivist framework. This is difficult because subjectivists have not been sufficiently forthcoming about what motivates their project such that we could then infer what moves would be compatible with it and what moves would not. Further, critics of subjectivism have been singularly unimaginative in their understanding of what motivates subjectivism and have assumed that only a very narrow range of moves are available to the subjectivist. At the beginning of this paper I mentioned (albeit too briefly) a few core

Christine Korsgaard called our attention to the need for subjectivists to make sense of processing failures, as well as potential troubles they may have in doing so, in two influential papers (Korsgaard 1996 and 1997). But Korsgaard chose to champion, at least in the more recent of these two papers, the most sweeping complaint one could make in this area. She argues that subjectivists cannot offer accounts of practical reason that can serve as guides that we might fail to follow. That is, she claimed that subjectivists cannot offer an account of our reasons that can tell us to do what we might not do.[12] She writes, "the empiricist [Humean] account explains how instrumental reason can motive us, but at the price of making it impossible to see how they could function as requirements or guides" (1997, p. 219). Korsgaard claims that Humeans lack the resources to make sense of a notion of good deliberation that could recommend an action that we might not choose. Thus her claim encompasses the claim that subjectivists can make no room for any errors in processing, but it is much broader still.

I think this claim of Korsgaard's is fairly obviously mistaken. But we should not let Korsgaard's extreme claim distract us from the important issue that her work helps us to see. The important question is whether or not the subjectivist can make sense of, and motivate a method of overcoming, all the kinds of errors of processing that there are. If there are kinds of errors that are possible in processing that subjectivists lack the resources to overcome, then subjective accounts still suffer a seemingly decisive objection. Korsgaard's claim was bolder than it needed to be to make real trouble for the subjectivist. Here I will investigate the merits of the more cautious, and more genuinely worrisome, version of the Korsgaardian objection against subjectivism.

The subjectivist project is to construct a vantage point for *P* from which a specified sort of concern determines *P*'s reasons. If, even from the subjectivist's preferred vantage point, it is possible that mistakes in processing occur in the generation of the specified sort of concern, then it is hard to see how the resulting concerns could track, let alone determine, *P*'s reasons. Thus our subjectivist seems forced to aspire to construct a vantage point from which such errors are impossible. The worry is that this cannot be pulled off within the subjectivist framework.

To make progress on this issue, we will need to get some handle on the kinds of errors in processing that an account of practical reason must

---

commitments that subjectivists share that could help structure the debate about whether this or that proposal is compatible with it.

[12] I take issue directly with Korsgaard's arguments for this claim in part two of my (2001b).

accept if it is to be adequate. Further, I want to explore the variety of ways that subjectivists have attempted to make room for such processing errors. Some subjectivists simply deny the problem, while others radically refashion their views in hopes of accommodating the possibility of such errors in processing. My attention will be focused on the question of whether or not subjective accounts are undermined because they are unable to offer an adequate response to the real distinction between good and bad processing. My main question is whether or not the subjectivist's resources are adequate to the job. Thus it is relevant both what resources we think the subjectivist has and how big the task is. Understanding how big the task is will be to understand what kinds of processing errors there can be. As I have said, I will conclude tentatively that it seems that the subjectivist's resources are adequate to the task.

4. Ideal Rationality, Not Ordinary Rationality

As has been said, the thought that an account of practical reason must make room for errors in processing can be addressed against a subjectivist account offered as a truth-maker of reasons claims. The thought would then be the subjectivist cannot offer a plausible account of the truth-maker in this area because their account of the truth-maker ignores the genuine possibility of processing errors. Thus it is likely to give wrong answers about the reasons people have. And if this is so, then subjectivism cannot be a good account of the truth-maker of reason claims.

This is to say that the subjectivist needs an account of good processing as part of their account of the truth-maker of reasons claims. But if this is what the subjectivist needs it would be a mistake to think of the needed account of good processing as simply an account of rationality, where rationality is conceived of as, among other things, a matter of making sensible uses of information in the real world. Let us call deliberation "ordinary rational" if, given time-constraints and sensible heuristics and biases that an agent might adopt to counter-act predicable patterns of human weakness in deliberation, the agent's deliberation in light of available information was good. Let us call deliberation "ideally rational" if it counts as perfect when we make no excuses for time-constraints or sensibly imposed heuristics and biases and such.[13]

[13] For a similar point see Railton (1986, p. 16), where he writes: "A fully informed and rational individual would, for example, have no use or desire for psychological strategies

What the subjectivist needs, according to the argument under consideration, is an account of ideal processing to go with her account of ideal information if she is to be able to offer a plausible account of the truth-maker of reason claims.[14] For if the topic of subjectivism is what an agent really has reason to do, and not merely what it makes sense for her to conclude that she has reason to do given her epistemic situation and time-constraints, then what the subjectivist needs is an account of ideal rationality or ideal processing.

Thus our subjectivist need not concern herself with cases in which a person processes sensibly given time constraints, uncertainty and such but fails to have a concern for that which she truly has a practical reason. If it is appropriate to call cases of this type of thing an error of practical reason, it is a kind of error that the subjectivist need not worry about in offering an account of the truth-maker of reason claims. Most obviously, subjectivists should remove considerations about the cost of deliberation and time pressures in arriving at a decision from playing a role in ideal processing. Understanding what the subjectivist's target notion needs to be helps us see which complaints are on topic and which are not.

5. Alleged Processing Errors and Subjectivist Replies

There are three general responses the subjectivist might have to a purported kind of error of practical processing, namely rejection, indifference, and attempted accommodation. Rejection involves the claim that the purported processing error is no real error at all. Alternatively, the subjectivist could allow that a purported error of processing is a real error, but claim that such errors are powerless to undermine the subjectivist's central claim and thus that subjectivists do not need to provide a fix for such errors. Call this response "indifference." Finally, the subjectivist might allow that the purported error is a genuine error and that, left uncorrected, such an error would undermine the subjectivist program, and thus attempt to find a subjectivist friendly fix for the kind of error involved. The remainder of this paper will consider various alleged possible errors of practical processing along with at least one of the three above subjectivist responses.

---

suited to circumstances of limited knowledge and rationality [. . .]."

[14] I assume throughout, unless otherwise mentioned, that the best subjectivist accounts will involve an idealized information component.

6. Weakness of Will

Our subjectivist might claim to be indifferent to some allegedly problematic cases. The most obvious case here would be weakness of will. Plausibly it does not matter for the truth of the subjectivist's theory whether weakness of will is possible or not. Of course it is possible, but our subjectivist should claim that this does not matter because weakness of will locates problems in choices rather than concerns and the subjectivist account makes use only of concerns, not choices, in the construction of her theory. Thus our subjectivist should claim that they can be indifferent to the possibility of errors of practical processing that are due to weakness of will because such errors, even if they are allowed to be possible, could not undermine the subjectivist account.

As I mentioned above, Christine Korsgaard is perhaps the contemporary philosopher who has most urgently pressed the subjectivist on the topic of providing an account of errors in processing. In (1996) she argues that "there is no reason to deny that human beings might be practically irrational in the sense that Hume considers impossible: that, even with the truth at our disposal, we might from one cause or another fail to be interested in the means to our ends" (p. 321). Rather, she claims, the Humean skeptic about practical reason "ought to allow for at least one form of true irrationality, namely, failure to be motivated by the consideration that the action is the means to your end" (p. 319). Let us call this error weakness of will. There is no critique here of the end, rather only a critique of the agent's failure effectively to pursue it. But then there can be no problem for the subjectivist account in this neighborhood. The subjectivist uses the agent's concerns to construct an account of the appropriate end for an agent. The subjectivist should say that the agent has a reason to take action to best achieve her ends as determined by her concerns. If the agent fails to do this, she has, by the subjectivist's lights, failed to act in accord with her reasons. Problems in acting in ways that best achieve our concerns will not alter the subjectivist account of what a person has reason to do.

Korsgaard, in the earlier paper, allows that the Humean, as opposed to Hume, has the resources to make room for such weakness of will. Yet, she claims that "once this kind of irrationality is allowed in the means/ends case, some of the grounds for skepticism about more ambitious forms of practical reasoning will seem less compelling" (p. 321). Her main point in invoking the possibility of true irrationality was to argue that a plausible account must say that our reasons motivate us insofar as we are rational, not that our reasons necessarily motivate us

even when we are irrational. The example of weakness of will is used to show that even knowing the facts a person might be irrational and so unmoved by her reasons. I have no quarrel with this claim (except I would say a person might count as rational and still be unmoved by her reasons because she might unculpably lack pertinent information). However, Korsgaard suggests that the subjectivist will have a hard time making room for weakness of will without opening the floodgates to non-subjectivist elements and it is this claim that I am resisting.

The subjectivist, of course, needs an account of which of an agent's concerns are authoritative and which are not. And, of course, it is contentious whether or not the subjectivist can adequately provide such an account. But this is not the objection here under consideration. For the problem of weakness of will only occurs after the authoritative ends have been identified. Thus, let it here be granted that the subjectivist has adequately made out this distinction. Thus imagine a case in which the authoritative concerns tell one to *X* but other, less authoritative concerns tell one to *Y*. Is it problematic for a subjectivist account that a person might knowingly choose an option at the urging of a less authoritative desire?

I think this cannot be problematic for the subjectivist. If the story is spelled out as I have it above, the subjectivist obviously suggests that it is the authoritative concerns that make it the case that one has most reason to *X*. To think that weakness of will is problematic for subjectivist accounts one must think either that subjectivists need to adopt the implausible "revealed" accounts of preference or concerns, in which what one chooses determines what one cares about or that it is impossible to care about an end without being motivated to take the means to that end. But subjectivists are far from being tied to such an account of concerns.[15] An agent's concerns might well determine her reasons even if weakness of will keeps her from acting in accord with those reasons. Because weakness of the will locates problems in choices rather than concerns it is powerless to constitute an objection to subjective accounts that claim it is an agent's concerns that ground her reasons. Genuinely problematic cases of processing errors for the subjectivist must occur before the appropriate end has been determined.

Even someone convinced by my claim tha weakness of will cannot create problems for subjectivist accounts of reasons for action might balk at my use of Korsgaard in this connection. After all, the conclusion of her argument is that subjectivists lack the resources to develop an account of

[15] Korsgaard does seem to claim that subjectivists are tied to a revealed account of a person's preferences. For compelling arguments against this view, see Hubin (2001).

rationality according to which people sometimes act irrationally. One could read Korsgaard two different ways. First, one could claim that her topic is not my notion of a reason at all, and her central claim involves problems that subjectivists have with rationality. I have addressed this reading elsewhere (Sobel 2001b, Part II, subsection C). Amongst the costs of this reading are that, I have argued, it has Korsgaard failing to address the views of Hume and Williams who are developing accounts of reasons. Secondly, one could see Korsgaard's argument to be addressed against a subjectivist account of the truth-maker of reason claims and claiming that there is a problem with the practical processing component of any such account. I am here reading Korsgaard in this latter way. On this view one would be right to see Korsgaard as addressing the question of ideal practical processing and claiming that problems for the subjectivist in this neck of the woods scuttles the subjectivist's account of reasons for action.

Gavin Lawrence argues differently to the conclusion that subjectivism founders in accounting for weakness of will.

> Thus over ends there is, for example, the practical irrationality of akrasia, viewed as the irrationality of agents' pursuing an end, whose efficient attainment they may calculate, but which they believe (truly or falsely) they ought not to be pursuing in the first place – since, they suppose, the good thing for them to do is something else. This description of akrasia essentially involves the idea of a kind of rational end assessment which the ER [End-Relative aka neo-Humean instrumentalism] conception rejects. Akrasia so characterized would then not be a possible practical irrationality on ER; indeed notoriously those ER theorists wishing to allow akrasia's existence have faced considerable problems in giving a description of it that both intuitively captures the phenomenon and preserves some sense of its practical *irrationality*. (Lawrence 1995, pp. 128-129)

Presumably the thought is that thinking to oneself that the good thing for me to do is this rather than that is already a thought that the subjectivist cannot make sense of. Lawrence seems to claim that thinking that the good thing for me to do is *X* rather than *Y* already brings with it the kind of rational end assessment that the subjectivist has rejected. But consider a person who believes a subjectivist theory, according to which her desires give her reasons. This is, of course, someone who thinks that subjectivism offers a good account of what they have reason to do. Such a person will develop views about what they ought to be doing (according to the theory) and occasionally feel tempted to do otherwise (unless the subjectivist view in question implausibly claims that it is the

desire that pushes her the strongest at the moment that is always the authoritative concern). Thus this person will think that there are things that they ought not to be pursuing. This person can think to herself that the wisest choice would be *A* and that she should stop being tempted by *B*. And this person could give in and choose *B*. Isn't this weakness of will seen as practical irrationality within a subjectivist framework?

Perhaps the subjectivist is thought to be unable to move from what I have so far claimed is compatible with the view to the thought that doing what the agent thinks is wise is "the good thing for me to do." But if this phrase does not receive a special philosophical gloss, it is just the sort of thing our convinced subjectivist might say about the option that is recommended by the view of practical reason she accepts.

Subjectivism is a theory of reason for action. It is compatible with the thought that reasons can have different weights and that one's reasons are not determined by one's current most oppressive craving. This, it seems to me, is all that is needed to make akrasia a possible way of being irrational within a subjectivist framework. Thus I have argued that

(1) subjectivists can make room for weakness of will and offer at least an initially plausibly explanation of it;
(2) the existence of weakness of will is powerless to constitute an objection to the philosophically popular variant of subjectivism that ties reasons to concerns rather than choices.

Thus I think weakness of will cannot be an example of poor practical processing that makes trouble for the subjectivist story of reasons for action. No doubt a complete practical theory should have something to say about rationality as well as reasons, and it would be a problem for the broadly Humean program if there could be no sensible account of rationality within such a framework. I do not see any reason to think that subjectivism is incompatible with a sensible account of rationality, but here I have only been at pains to point out that in any case, subjectivist accounts of reasons for action are safe from worries stemming from weakness of will.

7. Going Haywire

Korsgaard helpfully puts forward another line of thought to the effect that subjectivists cannot make sense of the difference between good and bad processing. She writes,

> But as Nagel points out in *The Possibility of Altruism*, the specifically rational character of going to the dentist to avert an unwanted toothache depends on how the belief and the desire are 'combined'. It is certainly not enough to say that they jointly *cause* the action, or that their bare co-presence effects a motive, for a person might be conditioned so that he responds in totally crazy ways to the co-presence of certain beliefs and desires. In Nagel's own example, a person has been conditioned so that whenever he wants a drink and believes the object before him is a pencil sharpener, he wants to put a coin into the pencil sharpener. Here the co-presence of belief and desire reliably lead to a certain action, but the action is a mad one. One might be tempted to say that a soda machine, unlike a pencil sharpener, is the source of a drink, so that the right kind of conceptual connection between the desire and the belief obtains. But so far that is only to note a fact about the relationship between the belief and the desire themselves, and that says nothing about the rationality of the *person* who is influenced by them. If the belief and the desire still operate on that person merely by having a certain causal efficacy when co-present, the rational action is only accidentally or externally different from the mad one. After all, a person may be conditioned to do the correct thing as well as the incorrect thing; but the correctness of what she is conditioned to do does not make *her* any more rational. (Korsgaard 1997, pp. 220-221)

Donald Hubin has offered considerations that put a similar kind of pressure on the subjectivist. Hubin points out that we can make sense of a bit of factual input into practical deliberation causing the agent to "go haywire." We could imagine that, as a causal matter, some bits of information produce wild results in a person's motivations. Just as some computer chips years ago were disposed to make wild calculations upon receipt of certain inputs, we can make sense of this as a possibility in agents as well.

Hubin supposes that the possibility of this sort of problem undermines only certain versions of subjectivism. He seems to think that such a possibility would undermine informed desire subjectivist accounts but would not undermine subjectivist accounts that fixed on "intrinsic actual" desires. I take it the thought is that the counterfactual deliberation might trigger some such instance of going haywire and so subjectivist accounts that look to our (counterfactually) informed desires must be able to distinguish what counts as going haywire from making sensible use of information. Hubin's story involves a tacit skepticism about the possibility of the subjectivist managing to mark such a distinction using an informed desire account (Hubin 1996). As Hubin sees it, his own actual intrinsic desire account has an advantage over counterfactual desire accounts in that the former need not worry about such glitches.

The reason he thinks his account is immune to such problems is that if a glitch caused one to have an actual intrinsic desire, Hubin is willing to say that this glitch has changed what one has reason to do (Hubin, conversation). But, obviously, the fact that I would have a glitch if I considered certain information, does not mean that these counterfactual glitches affect what my reasons are prior to the glitch.

Hubin's actual intrinsic desire account suggests a subjectivism that might have some advantages over informed desire accounts.[16] Thus, his account might offer help to the subjectivist in making clear how they can explain and accommodate the possibility of poor processing. Yet I am tempted to think that if Hubin's "haywire" objection were telling against informed desire versions of subjectivism, that it would likely be telling against his intrinsic actual account since, I suspect, facts about what a person would want in non-actual circumstances will play a role in determining what a person actually intrinsically wants. And if so, then the worry about counterfactual glitches will reoccur. I want to not, as much as possible, take a stand about what the best version of subjectivism looks like. But so far I do not see how the glitch point could doom informed desire accounts without dooming actual intrinsic desire accounts as well.

Richard Arneson explicitly picks up Hubin's complaint and finds it sufficiently forceful to justify abandoning subjectivism. Arneson, in discussing subjectivist theories of well-being, writes:

> It might simply be a brute psychological fact about me that if I were to become fully informed about grapes, this process would set off a chemical process in my brain that would lead me to crave counting blades of grass on courthouse lawns as my primary life aim. This would seem to

[16] Hubin, as I understand him, is tempted by expressivism about all things considered judgments about reason claims. A person could attempt to combine subjectivism and expressivism is different ways. One model would mimic the combination of utilitarianism with expressivism about morality. Here the idea would be, I take it, that one's normative ethical view was utilitarian but one was a meta-ethical expressivist. Absent tricky moves, on such a view one would admit, when wearing one's meta-ethical hat, that none of one's normative-ethical claims was literally true or false. Alternatively, one could combine the two by having subjectivism speak to a subset of the concerns that feed into an all things considered judgment about what there is most reason to do, but be an expressivist, or at any rate non-cognitivist, about judgements of how to combine such subsets into all things considered judgements about what there is reason to do. I believe that both David Copp and Hubin hold this latter sort of view. Copp's "needs and values principle of self-grounded reasons" tells a quasi-subjectivist story about what he calls "self-grounded reasons" but Copp denies that there are facts about what one has reason to do all things considered, and even denies that the question of what one has reason to do all things considered is in good order. See Copp (1997, pp. 86-106).

> be an oddity of my brain, not an indicator of my true well-being. (Arneson 1999, p. 134)

Many of the examples that people offer of glitches are cases in which either the outcome of deliberation seems crazy or there seems to be a lack of relevance between the consideration and the resulting concern. Why should it be thought that subjectivists can not make room for a notion of relevance of a consideration to a concern and suggest that the authoritative concerns are those that are arrived at only by relevant considerations? Offhand, it seems continuous with subjectivism to say that it is only desires that arise as a result of appreciation of the object that carry authority. Desires for a life of grass counting that arise causally from consideration of grapes need not count for the subjectivist as authoritative. This would involve developing a subjectivist friendly notion of relevance.[17] Now as far as I know no one, neither subjectivists or objectivists, have a worked out notion of relevance that could underwrite the distinction between glitchy and non-glitchy processing. My goal for now is simply to point out that some notions of relevance could seem subjectivist friendly.

Note that crucially this notion of relevance would have to be what I will call procedural rather than output driven. That is, it does not presuppose that certain objects or options are more intrinsically worthy of being desired than others and then count deliberation as rational only if it leads to those objects. Rather, the thought here is neutral with respect to what can sensibly be desired after this process. Similarly a typical subjectivist requirement that one's desires be transitive is procedural rather than substantive. Such requirements put constraints on coherent patterns of concerns, but do not rule out patterns because they involve this or that element. Procedural requirements on authoritative desires can, of course, be implausible. Consider the procedural requirement that all and only desires on Tuesdays have authority. And of course there can be implausible versions of subjectivism. The thought here is just that procedural requirements on authoritative desires are compatible with a thoroughgoing subjectivism.

The subjectivist denies that an agent's reaction counts as a glitch merely because it is for a strange thing. If the objection under consideration were merely that subjectivists sometimes tolerate the thought that people have reason to go in for strange things then all this talk about glitches would be unnecessary. That is, if glitches are

[17] As will become clear, I do not aspire to offer here such an account. I merely mean to question why the possibility of such an account might feel closed to a subjectivist.

attributed retrospectively upon seeing what the deliberation hit on, then it is not the thought of a glitch that drives the resistance to subjectivism, but rather the thought that certain ends are rationally mandatory (or at least that certain ends are rationally forbidden). So I will ignore retrospectively attributed glitches as they just follow from the rejection of subjectivism rather than motivating it.

How then could we best capture an intrinsic rather than derivative or retrospective notion of a glitch? That is, what would make a bit of processing count as intrinsically glitchy? I think this is a harder question than it looks. Again, my main concern is to suggest that the subjectivist is in no worse of a position to acknowledge whatever genuine glitchyness that might exist. So the task of a person who wants to use the notion of a glitch against the subjectivist must be to bring forth cases that genuinely count as glitches and show that subjectivists lack the resources to so categorize them.

Here is one possibility. Suppose that there are proper ways for the brain to function and glitchy ways for the brain to function and that in principle brain scientists can tell us whether or not a particular deliberative process involved a brain glitch. If this is so, and if the scientists categorize a glitch procedurally, based not on the output of desire but on the causal process itself, then presumably the subjectivist could simply accept that it is the desires that would be produced under the appropriate conditions where no such glitches are involved. Such a notion of a glitch brings with it no presuppositions about what a person must want in order to be glitch free. If the scientist categorized glitches based on outputs of deliberation we would again be back to the starting point of the dispute between subjectivists and objectivists, not in possession of a powerful argument for objectivism. Thus, as far as I can see, the possibility of glitches of the sort our objectors have mentioned would not undermine the plausibility of subjectivism. Procedural glitches are compatible with subjectivism and output glitches have not yet been shown to offer obvious cases of problematic processing.

I take it that Kantians typically want to have a procedural account of good processing as well since they want it to be the case that practical reasoning does not find normative facts but creates them. We are free insofar as we guide ourselves by giving ourselves laws rather than having laws imposed from without. But if this is so, Kantians need a procedural account of good processing just as subjectivists do. So accommodating glitches within a procedural account of good processing is a task for subjectivists and Kantians alike. As far as I see, we have no arguments before us that objectivists are better positioned to be able to capture the

wanted notion of an intrinsic glitch. And attributing a glitch where deliberation hits on things that are held to be necessarily irrational simply begs the question against subjective accounts.

8. Acculturated Desires

A familiar criticism of subjectivism is that it founders because what we desire is a function of what we are used to, think plausible, or already believe good. Thus the person brought up to believe that she is not worthy of a vote or not worthy of being touched by "better" types of people might not form desires to vote or be seen and touched without shame. She might "not dare" to desire certain things that she has been taught are not appropriate to someone such as herself. Thus the desires of such ill used people will reflect their horrendous upbringing and not reflect what such people really have reason to do. This, I suppose, might be said to be one sort of common processing error.

Subjectivists have a variety of responses to this objection. First, and perhaps most importantly, typically part of the idealization process that subjectivists recommend is having one's false beliefs discredited and being confronted with true beliefs. Now typically the kind of situation I have described will be rife with factually erroneous beliefs.

Further, typically in such cases our worry is that an agent might lack sufficient familiarity with various options or not feel it their place to investigate certain sorts of options. The subjectivist account should suggest that it is only after an agent has been exposed to the full variety of ways that she might live that her concerns are authoritative.[18] Thus worries that she might have crimped desires due to lack of appreciation of the wonders of the options that she has not come across in real life, or worries about how she might not feel it her place to experience some ways of life, are simply out of place.

Of course it can happen that as a result of a horrendous upbringing an agent lacks reason to pursue that which she would have had a reason to pursue but for the upbringing. Such experiences can change what we have reason to do. If those options that she would have had a reason to pursue no longer resonate with her when she gives them a fair trial, she might well have lost her reason to go in for such things. It is a strength of the subjectivist story that it allows for the possibility that severely damaged people might have importantly different reasons than the rest of

[18] There are, I think, real problems with the typical "full information" account of the authoritative vantage point. See Velleman (1988), my (1994), and Rosati (1995).

us. Further, even if the subjectivist story can explain why a particular person has a reason to be able to be seen in public without shame, it may yet be that such an agent has most reason to keep her head down and not fight the system so as to avoid being pummeled. Or she may not, depending on her concerns.

9. Failures to Live Up to the Aristotelian Virtues

The subjectivist might reject some kinds of alleged errors of practical processing as not genuinely errors at all. Undoubtedly this would be the typical subjectivist's attitude towards the claim that the Aristotelian virtues are excellences of practical processing, such that one makes a practical error if one fails to act virtuously. If this were true, then subjectivism would have to be mistaken. Our subjectivist will no doubt claim that virtues such as generosity or bravery, while they may typically tend to make others love us (which almost all in fact care about) or tend to help us achieve our projects whatever they may be, nonetheless have no necessary connection to excellence in practical processing.[19] This rejection will need defense, as will the claim that the rejection here is implausible.[20] The important question is whether or not we have here a clear case of poor processing. The issue is whether the claim that failures of virtue are necessarily failures of ideal rationality is clearly enough true that we can use this fact to shape our theory of reasons to fit it or if it is a contentious claim that simply counts as the rejection of subjectivism rather than a clear rationale for becoming a non-subjectivist. Bluntly, I do not think this claim is secure enough to serve as an Archemedian leverage point in this debate. This is not an argument that there could not

[19] The example of bravery might suggest an alternative subjectivist strategy that I will not discuss. Our subjectivist might try to "appropriate" a given alleged case in which practical processing can go well or ill in ways that render it compatible with a subjectivist story. So our subjectivist might say that bravery is an excellence of practical processing but understand bravery in a way such that it, at least when combined with an ideal theoretical reason, necessarily helps one along in the accomplishment of one's concerns. Philippa Foot, in (1978), suggests a picture of some of the virtues that looks somewhat like this. She suggests that the virtues are names for ways of avoiding common errors in acting according to reason. So bravery is a virtue because we tend to fall away from what reason (independently) dictates when reason requires actions that are fearful. On this picture the virtues do not fix what one has reason to do, rather they are generally useful for getting us to do what reason has already picked out as the appropriate aim.

[20] The most thoughtful attempt that I know of to make out the case that failures of virtue are necessarily failures of ideal rationality is in Quinn (1993) and Hursthouse (1999, Part 3).

be an argument showing that virtue failures are failures of ideal rationality. It is merely an argument that absent such an argument, the claim under consideration is clearly not decisive.

10. Depression

The topic of depression and whether to think of it as a possible cause of errors of processing is interesting and I will be unable to address it adequately here. I will confine myself to two claims that seem fairly clear in this neighborhood.

First, the extent to which depression will be seen as a source of practical processing errors will likely hinge somewhat on what the subjectivist take the relevant attitude to be. If the relevant authority conferring attitude were thought to be a desiring, then it seems likely that depression could dampen and perhaps extinguish such concerns. On the other hand, if the relevant attitude were thought to be valuing, then arguably depressions ability to dampen or perhaps deaden desire need not be thought to necessarily affect the agent's valuations. Michael Smith, for example, argues that valuing is compatible with having no motivational desire for the object of positive valuation. Thus, depression and the like might have less ability to produce errors of processing if the relevant attitude is one of valuing.

Second, attitudes such as depression cause fairly systematic loss of affect and lively concern. A subjectivist could claim that attitudes that dampen or extinguish the relevant concern, without regard to its object, could be held to be pathologies that must be extinguished or circumvented if we are to get at the relevant authoritative concerns. That is, again the subjectivist could look to make a procedural case against depression rather than a outcome driven argument that it counts as or leads to, errors of practical processing.

11. Conclusion

The main question here is not the viability of this or that particular version of an informed desire account of reasons. The main question is whether we find kinds of errors in processing that subjectivists are in a worse position to acknowledge and accommodate than non-subjectivists. What we are looking for is a plausible account of what goes wrong when errors in processing occur that subjectivists cannot help themselves to. So

it seems we would need a clear example of a processing (not merely outcome) error and then have a case made that subjectivists are unable to count this as an error of processing. I don't see that we have candidates yet that purport to show all this.

Acknowledgements

I am grateful to Andrew Altman, David Copp, Justin D'Arms, Janice Dowell, Claire Finkelstein, Don Hubin, Doug Lavin, Stephen Perry, Connie Rosati, Sergio Tenenbaum, and Kit Wellmon for helpful feedback about the issues raised in this paper. I presented this paper to The University of Pennsylvania's Legal Theory Workshop, The Ohio Reading Group in Ethics (ORGiE), and Georgia State University. I am very grateful for the beneficial discussions that ensued.

Bowling Green State University
Department of Philosophy
Bowling Green, OH 43403
e-mail: david_sobel@hotmail.com

REFERENCES

Arneson, R. (1999). Human Flourishing versus Desire Satisfaction. *Social Philosophy and Policy* **16** (1), 113-142.

Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford: Clarendon Press.

Copp D., and D. Sobel (2002). Desires, Motives, and Reasons: Scanlon's Rationalistic Moral Psychology. *Social Theory and Practice* **28** (2), 243-276.

Copp, D. (1997). The Ring of Gyges: Overridingness and the Unity of Reason. *Social Philosophy and Policy* **14** (1), 86-106.

Cullity, G. and B. Gaut, eds. (1997). *Ethics and Practical Reason*. New York: Oxford University Press.

Darwall, S. (1983). *Impartial Reason*. Ithaca, NJ: Cornell University Press.

Foot, P. (1978). Virtues and Vices. In: *Virtues and Vices*, pp. 1-18. Berkeley, CA: University of California Press.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Clarendon Press.

Griffin, J. (1986). *Well-Being*. Oxford: Oxford University Press.

Hare, R.M. (1981). *Moral Thinking: Its Levels, Methods and Point*. Oxford: Clarendon Press.

Harsanyi, J.C. (1982). Morality and the Theory of Rational Behavior. In: A.K. Sen and B. Williams (eds.), *Utilitarianism and Beyond*, pp. 39-53. Cambridge: Cambridge University Press.

Hubin, D. (1996). Hypothetical Motivation. *Nous* **30** (1), 31-54.

Hubin, D. (2001). The Groundless Normativity of Instrumental Reason. *The Journal of Philosophy* **98** (9), 445-468.

Hume, D. (1967). *A Treatise of Human Nature*. Edited by L.A. Selby-Bigge. Oxford: Oxford University Press.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford: Oxford University Press.

Johnson, R. (1999). Internal Reasons and the Conditional Fallacy. *Philosophical Quarterly* **49** (194), 53-71.

Kagan, S. (1989). *The Limits of Morality*. Oxford: Oxford University Press.

Korsgaard, C.M. (1996). Skepticism about Practical Reason. In: *Creating the Kingdom of Ends*, pp. 311-334. Cambridge: Cambridge University Press.

Korsgaard, C.M. (1997). The Normativity of Instrumental Reason. In: Cullity and Gaut (1997), pp. 215-254.

Lawrence, G. (1995). The Rationality of Morality. In: R. Hursthouse, G. Lawrence and W. Quinn (eds.), *Virtues and Reasons: Philippa Foot and Moral Theory*, pp. 89-148. Oxford: Clarendon Press.

Lewis, D. (1989). Dispositional Theories of Value. *The Aristotelian Society Supplementary Volume* **63**, 113-137.

Quinn, W. (1993). Rationality and the Human Good. In: *Morality and Action*, pp. 210-227. Cambridge: Cambridge University Press..

Railton, P. (1986). Facts and Values. *Philosophical Topics* **14**, 5-31.

Railton, P. (1997). On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action. In: Cullity and Gaut (1997), pp. 53-80.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rosati, C. (1995). Persons, Perspectives and Full Information Accounts of a Person's Good. *Ethics* **105**, 296-325.

Rosati, C. (1996). Internalism and the Good for a Person. *Ethics* **106**, 247-273.

Scanlon, T.M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T.M. (2002). Replies. *Social Theory and Practice* **28** (2), 337-342.

Seanor, D. and N. Fotion, eds. (1990). *Hare and Critics: Essays on* Moral Thinking. Oxford: Clarendon Press.

Sidgwick, H. (1981). *The Methods of Ethics*, 7th edition. Indianapolis, IN: Hackett.

Smith, M. (1994). *The Moral Problem.* Oxford: Blackwell Publishers.

Sobel, D. (1994). Full Information Accounts of Well-Being. *Ethics* **104**, 784-810.

Sobel, D. (2001a). Explanation, Internalism, and Reasons for Action. *Social Philosophy and Policy* **18** (2), 218-235.

Sobel, D. (2001b). Subjective Accounts of Reasons for Action. *Ethics* **101**, 461-492.

Sobel, D. and D. Copp (2001). Against Direction of Fit Accounts of Belief and Desire. *Analysis* **61** (1), 44-53.

Velleman, D. (1988). Brandt's Definition of 'Good'. *The Philosophical Review* **97**, 353-371.

Williams, B. (1981). Internal and External Reasons. In: *Moral Luck: Philosophical Papers 1973-1980*, pp. 101-113. Cambridge: Cambridge University Press.

Williams, B. (1995). Internal Reasons and the Obscurity of Blame. In: *Making Sense of Humanity: and Other Philosophical Papers 1982-1993*, pp. 35-45. Cambridge: Cambridge University Press.

Sergio Tenenbaum

THE CONCLUSION OF PRACTICAL REASON

Aristotle's famous contention that the conclusion of practical reasoning is an action (see Aristotle 1984) (henceforth "the Aristotelian Thesis") often baffles action theorists. I will first examine a few reasons to object to the Aristotelian Thesis; these objections seem to support the view that the conclusion of practical reasoning is an *intention*. However, I will argue that this is not a tenable position, and I will propose a way to understand the Aristotelian Thesis that can overcome these objections. The first part of the paper examines the case against the Aristotelian Thesis, and in favour of the main alternative view, the view that the conclusion of practical reasoning is an intention. It seems that the Aristotelian is vulnerable to a few rather obvious objections, while the alternative view seems to face none of these difficulties. The following sections try to show that appearances here are wholly deceptive. In the second section, I argue that when we properly understand the subject matter of the Aristotelian Thesis, that is, when we understand what can be properly considered a *conclusion* of practical reason, it turns out that the alternative view is indefensible. The third section argues that, on the other hand, with this proper understanding of its subject matter in hand, we can show that the Aristotelian Thesis is immune to the objections canvassed in the first section.

Before we move on, a piece of terminology is helpful. The conclusion of practical reasoning is not a prediction about how one will act or intend (see, on this issue, Korsgaard 1997). This can be registered by adding a 'should' to the conclusion ("I *should* turn the light on"). It is my view, however, that if the Aristotelian Thesis is correct, this way of writing the conclusion is a categorical mistake. So in order not to prejudge the issue, I will write the conclusion of practical reasoning as an intention or action, depending on the view being discussed, but maintain throughout that good practical reasoning (or a sound argument in the practical

In: Sergio Tenenbaum (ed.), *Moral Psychology* (*Poznań Studies in the Philosophy of the Sciences and the Humanities*, vol. 94), pp. 323-343. Amsterdam/New York, NY: Rodopi, 2007.

sphere) *justifies*, rather than guarantees the truth (or the reality) of the intention or the action.

1. The Case Against the Aristotelian Thesis

The Aristotelian Thesis seems to be vulnerable to the following objections:

(1) An agent intends an action only under a particular description.
(2) Whether we succeed in carrying out a particular action does not depend solely on our reasoning capacities.
(3) Sometimes one does not act on the conclusion of practical reasoning (Robert Audi raises a similar objection in 1989, p. 93).

Let us start with (1). Actions have an "accordion effect" (this term was coined by Joel Feinberg; see his 1965). To use Davidson's well-known example, by flipping the switch, Mary may turn on the light, and also inadvertently alert a prowler (Davidson 1980). In this case, all these actions might be identical:

(a) Mary's flipping the switch.
(b) Mary's turning the light on.
(c) Mary's alerting the prowler.

It is somewhat controversial how actions should be individuated, and indeed whether the actions described by (a), (b) and (c) are actually one and the same action, or whether a narrower criterion for action individuation is more adequate (for a survey of the available positions, see Mele 1992). Since the objection is far more powerful if we assume that actions are individuated in this broad manner, I will assume that, if "*A* -ed" and "*A* -ed" are action descriptions, and that if "*A* -ed by -ing," then "*A* -ed" and "*A* -ed" refer to the same action. With this assumption in place, there seems to be an obvious problem with the Aristotelian Thesis. For certainly a sound piece of practical reasoning is made unsound by substituting descriptions of the same action.

Let us take, for instance, the following reasoning:

I. I want to (intend to, shall) read in the bedroom.
II. I believe that in order to read in the bedroom, I must turn the light on.

Therefore,

III. I turn the light on.

This seems like a valid piece of practical reasoning. One might dispute whether, for instance, premise (I) should not be substituted by an evaluation (such as 'is desirable' or 'is good'; see Scanlon 1998), or, of course if premise (II) should not state a fact as opposed to a belief, (see Thomson 2001, Ch. 2), etc. But it seems that some revised version of the reasoning above should turn out to be valid, at least if we leave aside for the moment the idea that any such piece of reasoning can count as valid only on the assumption of some kind of *ceteris paribus* clause, or that considerations that counted against turning the light on did not provide reasons not to perform the action that were strong enough to override the above reasoning.[1]

But if by turning the light on, I alert a prowler, it seems that (III) could be substituted by the following:

IV. I alert a prowler.

However, it seems that an argument that concluded (IV) from (I) and (II) would be invalid. One might object that the substitution in question is not allowed by the Aristotelian Thesis. After all, substituting co-referring expressions doesn't necessarily preserve the validity of an argument or a piece of reasoning. I doubt, however, that this move can work. For the Aristotelian Thesis does not claim that the conclusion of practical reasoning is the statement of an action, or a proposition that describes a certain action. It claims that it is the *action itself*. In theoretical reasoning substituting different statements of the same proposition[2] in an argument should not make a difference to its validity; similarly, we should also expect that substituting descriptions of identical actions should not make a difference to the validity of a piece of a practical reasoning if we accept the Aristotelian Thesis. One could quibble further, but I will assume that the Aristotelian Thesis is committed to allowing such substitutions.[3]

It might be worth examining a defence of the Aristotelian Thesis that concedes some of these points. One might think that although there is nothing wrong about having a sentence of the form "I intend to . . ." as the conclusion of practical reason, it would be wrong to deny the Aristotelian Thesis on this basis. Although I can move from a belief to

[1] Although we are leaving this aside for the moment, that the argument needs such assumptions will be a central concern of this paper.

[2] Talk of "propositions" in this way often sets off a number of warning bells. But since one's views on the nature or existence of propositions should not affect the point here, I will ignore them.

[3] Since my aim is to defend the Aristotelian Thesis, if this assumption turns out to be false, it would only bolster my case.

another by (theoretical) reasoning, it would be wrong to say that the conclusion of my reasoning is a *belief*, rather than its *content*. Similarly, although I form the intention on the basis of reasoning, it would be wrong to say that the conclusion of the reasoning is my intention rather than its content. And the content of the intention is an action (see Clark 1997, especially pp. 19-20). However, this line of reasoning is problematic in a few ways. First, what we get is a rather different version of the Aristotelian Thesis. On this reading, the Aristotelian Thesis turns out to be a claim about what is represented in the conclusion of a practical reasoning. The more contentious Aristotelian Thesis does not merely present a contrast between *what is represented* in the conclusion of theoretical and practical reasoning (the latter must be the representation of an action, former a representation of anything), but states that the latter is not a matter of representing, but of doing something. In Aristotle's words "whenever one thinks that every man ought to walk, and that one is a man oneself, *straightaway one walks*" (Aristotle 1984, 701a8-12, emphasis added). Moreover, if we reject the more radical Aristotelian Thesis and think of the conclusion in terms of the content of the intention of someone who reasons properly from (I) and (II), it is not clear that we should then accept that the conclusion of a practical syllogism is an action, rather than something *that is brought about through the action*, and in fact, a particular *description* of something that is brought about through the action (for a similar point, see Hornsby 1998, pp. 88-89). The content of the intention is not the action of my turning the light on, but rather "that the light is turned on," an effect brought about by my action. If one accepts Davidson's individuation of actions, there will be many effects for each action, and thus one cannot identify the action with what it brings about. Of course, one might take issue with Davidson's individuation of actions. But it is worth noting that one would need a rather fine-grained individuation of actions. For one would need a different action not only for each different effect, but a different action for each different possible object of intention. Assuming that Mrs. Jones is the prowler, "alerting the prowler" and "alerting Mrs. Jones" describe the same effect, but they are two different possible objects of intention (that is, one could intend to alert Mrs. Jones, but not intend to alert the prowler).

Let us now turn to objection (2). Suppose I go through what seems to be the same piece of deliberation, but instead of turning the light on, I inadvertently flip the alarm switch on. But now since I did not perform the action described in (III), it seems that, if the conclusion of a piece of

practical reasoning is an action, the only conclusion available to me is the following:

V. I turn the alarm on.

Concluding (V) from (I) and (II) does not speak very highly of one's intellectual powers. Yet, whatever my general limitations are, what I am displaying in this particular case is clumsiness rather than stupidity. One could reply that, in the absence of action, I should be regarded as concluding nothing at all. But this seems implausible; after all, it seems that I set myself to act because of the conclusion of a piece of reasoning. Alternatively, one could say that the action that concluded my practical reasoning should not be described as turning the alarm on, but rather as:

VI. I flip the switch.

However, even if we ignore the difficulties raised by objection (1), this cannot take us very far. If I am clumsy enough, I can also fail to flip the switch after deliberating in this manner. It seems that the only retreat that might accomplish anything, at least under certain views of the nature of mental states,[4] is to stop at a mental action that might be completely under the control of the agent, such as forming an intention or making a decision.[5] But this grants the opposition their point, since those who deny that the conclusion of practical reasoning is an action will likely argue that the conclusion of practical reasoning is the formation of a certain mental state (such as an intention, decision, etc.).

A final reply is open to the "Aristotelian." She can say at this point that the conclusion of the practical reasoning at this point is the following:

VII. I try to turn the light on.

For any such case, we could characterize the action as one of trying. It will turn out that this suggestion is essentially correct, but as it stands it seems problematic. One might now want to say that if (VII) is the only thing that (I) and (II) and my logical skills can guarantee, then the conclusion of any piece of practical reasoning must be a case of "trying."

---

[4] This qualification is necessary, since in some views there might be no "non-disjunctive" relevant mental states that are fully in control of the agent. For a view that suggests this possibility, see Hornsby (1998).

[5] Taking those mental states to be actions is itself problematic since it seems to generate an infinite regress. That is, if forming intentions are actions, then one should form intentions intentionally, and, under a plausible view of action, that means, that under certain description of this action, one intended to form an intention. But this means that one must have formed the intention to form the first intention.

But this suggestion faces a serious problem. If trying is anything that is not fully under my control, such as a bodily movement, even the most rational agent could fail to move from (I) and (II) to (VII) by sheer bad luck. If it is just a mental action or a mental event of some kind, then it seems that we are getting perilously close to the view that the conclusion is just an intention.[6] And this leads us directly to (c). For we would need to stretch the notion of "trying" pretty thin in order to cover all cases of practical reasoning. For suppose Larry wants to go for dinner this evening. Now it is 11:45, and he knows that he needs to call the restaurant at 12:00 to make reservations. He forms the intention to call at 12:00. Here it seems that Larry is done with the practical reasoning, and even if he were to die at 11:55, this would not change the fact that he had successfully carried out a piece of practical reasoning. His death would have prevented the action, but not the reasoning. Thus the conclusion of practical reasoning is not an action.

None of these problems seem to arise if instead of (III) we have, as the conclusion of (I)-(II):

VIII. I intend to turn the light on.

We cannot substitute "alert the prowler" for "turn the light on" in the context of an intention, and intentions can certainly fail to be carried out by clumsiness, death, etc.[7] The case against the Aristotelian Thesis seems compelling.[8]

---

[6] There is also another problem. It seems that what I should aim as the result of practical reasoning is the action itself, not an attempt. Although it is true that any time we *X*, we thereby also try to *X*, it is unclear that having *X*-ing and the attempt to *X* as one's aim amount to the same thing. If aim to hit a good serve, I will be concerned with the existence of obstacles, and I will try to remove them (I will, for instance, wait for the wind to stop blowing). But if I am concerned only to try to hit a good serve, I will not care about any external obstacles to my serve going in (no matter how hard the wind blows, my attempt will be flawless).

[7] I will assume that intentions are not actions. Although this makes for simpler presentation, rejecting this assumption will still leave room for a distinction between the Aristotelian view and its main alternative. The views would be distinguished as a different between *which* action should be regarded as the conclusion of practical reasoning: the intention to *X*, or *carrying out* the intention to *X*.

[8] The same advantages can be claimed for the view that the conclusion of practical reasoning is a practical judgment. See, for instance, Audi (1989).

2. Practical Reasoning and Soundness

Before we move on, we need to say a few words on what counts as valid and sound instances of practical reasoning. Of course, I do not mean here to give a complete account of the issue, but just place some constraints on these notions that will be relevant for the argument in this paper.

First let us start with the premises. We can distinguish between two kinds of premises that appear in practical reasoning. There will be premises that are possible contents of a belief clause, and whose soundness depends on whether the premise is true. On the other hand, there will be premises that specify a certain end or aim of the agent. The soundness of these premises will depend on whether the end is appropriate, correct, or one that the agent should have. I hope these remarks are neutral with respect to various views about the nature of practical reasoning. So, for instance, I do not assume that an instance of practical reasoning needs to have premises of the second kind. So, someone who thinks that desires, preferences, etc. are irrelevant to practical reasoning, will think that there are no premises of the second kind, and, perhaps that each premise describes the content of a belief of the agent (though some of the beliefs will be evaluative beliefs) or that all the premises are fact-stating propositions (though some of them state evaluative facts). A "subjectivist" about practical reason might think that all premises of the second kind are sound as long as they specify the content of the agent strongest desire, or the content of his relevant preferences, etc.

I will define an "acceptable conclusion" as follows: A conclusion is acceptable if and only if it is a conclusion that a fully rational agent would, or at least could, accept if she were in a relevantly similar situation. What counts as relevantly similar will vary from theory to theory, but it will probably involve having similar beliefs and preferences, that the agent be under similar time constraints, etc. The central idea is that an agent who is committed to an unacceptable conclusion is, on that account, irrational. I will assume that the following are unacceptable conclusions: a conclusion in which (or a conclusion that recommends that) one knowingly chooses a less preferred over a more preferred option, or a conclusion in which one takes (or a conclusion that recommends taking) what one knows to be insufficient means to one's ends. Depending on the details of one's theory of practical reason, different conclusions would count as unacceptable; I hope to use for my purposes only relatively uncontroversial examples of unacceptable conclusions. Similarly, a "fully rational agent" is one who is never guilty

of any kind of irrationality, but what counts as such an agent depends to some extent on one's theory of practical reason: here too, I hope to steer away from controversial examples.[9] Drawing an acceptable conclusion does not guarantee that the agent will do well, or even as well as possible for an agent in such a situation, for the agent might, for instance, lack important information.

We can now state constraints that will play an important role in the argument:

(C_1) A valid piece of reasoning never leads from sound premises to an unacceptable conclusion.

(C_2) If an agent acts irrationally then one of the following must be the case:
- (a) The agent forms noninferential beliefs irrationally.
- (b) The agent forms noninferential aims or ends irrationally.
- (c) The agent performs an invalid piece of reasoning.
- (d) Another irrational process that can be attributed to the agent takes place.

(C_1) is a relatively weak requirement on good (or valid) practical reasoning. Good practical reasoning should not move us from true beliefs and appropriate aims into a position in which we are guilty of some form of irrationality. This a rather basic constraint on good reasoning in general: that it should not take us from "unimpeachable" starting points into an irrational stance.[10] Indeed one could argue that a stronger constraint is also quite plausible: good reasoning shouldn't take us from *acceptable* premises into unacceptable conclusions. However, since I only need the weaker claim, I will commit myself only to (C_1). (C_1) is also formulated in such a way as to be neutral between whether the

[9] Of course a general sceptic about practical reason will not accept anything as an example of practical irrationality. But it is not clear to me that for such a sceptic practical *reasoning* is possible; general scepticism about practical reason would probably render this issue moot.

[10] This is not to say that all good reasoning is truth-preserving; no doubt, inductive reasoning is not like that. It also does not mean that we cannot end up with an unacceptable conclusion as the result of good reasoning; one would need then to conclude that some of one's premises were unacceptable, and, *ceteris paribus*, one would consider revising at least one of them. See Harman (1986). Harman does not endorse the use of notions such as "valid" for reasoning (since they seem to confuse reasoning and argument), but as long as one keep the distinction between reasoning and argument clear, I don't see any reason not to use the word 'valid' for reasoning that is conducted solely in accordance with correct or appropriate rules of inferences. However, substituting 'good' for 'valid' would not alter in any way the argument of the paper.

conclusion is an action or an intention. Certainly rational agents both intend and act, and they could both intend and act rationally or irrationally.

(C_2) simply tells us that if an agent acts irrationally, then the agent must be guilty of some specific failure of rationality. The general idea of (C_2) is that there could be no "blameless" irrationality.[11] The failure of rationality might be due to an irrational starting-point or to irrational inferences, but as long as the agent forms all cognitive states as a rational agent would, the agent could not be guilty of irrationality. Clause (d) just covers the possibility that we overlooked a cognitive failure that cannot be assimilated to (a)-(c); I leave it there since the arguments presented for the Aristotelian thesis do not depend on (a)-(c) covering all possible instances of irrationality.

I will also make a couple of assumptions: As I said above, I treat the following as cases of irrationality: (a) the agent knowingly acts counter-preferentially; (b) for which the agent knowingly pursues ineffective means to her ends. Although someone might find instances of (a) and (b) which it might be contentious whether the agent behaved irrationally, the examples I use are fairly straightforward. Also for the sake of convenience, I will assume that the agents we discuss have preferences similar to those that we expect that most agents have (they prefer more over less money; they prefer not to destroy their property, etc.), and that these preferences are in no way irrational.

Finally, it is also worth adding that although I assume that we can *attribute* instances of practical reasoning to agents, I am making no assumptions about the ontological commitments of this attribution. Perhaps each step in the agent's practical reasoning must reach consciousness, or perhaps each must have *some* kind of psychological or physical reality, even if they do not reach consciousness; perhaps, the attributions are fully determined by questions about how to best interpret the behaviour of the agent as rational, etc. My only commitment is to the *possibility* of attributing these instances of practical reasoning to the agent.

[11] I am not going to argue for this claim, so it is open to the "anti-Aristotelian" to argue against this claim. However it would be surprising to conclude that the anti-Aristotelian position is hostage to the possibility of blameless irrationality; those who reject AT typically do not reject it on these grounds.

3. The Case against the View that the Conclusion of Practical Reasoning Is an Intention

Let us now look more carefully at the Aristotelian Thesis. Take the following pieces of reasoning:

IX. No honest person becomes a millionaire just by chatting.
X. I am an honest person.
XI. If I intend to become a millionaire, I should do something other than chatting.

XII. If something is good in some respect, then I have some reason to pursue it.
XIII. Every beautiful thing is good in some respect.
XIV. Thus, I have some reason to pursue anything that is beautiful.

Now these are valid pieces of reasoning, and let us assume that all the premises are sound. Certainly these arguments do not have an action as a conclusion. They also seem to be species of practical reasoning. So do not we have here a fast refutation of the Aristotelian Thesis? If the Aristotelian Thesis is plausible at all, we need to restrict its scope. In particular, the Aristotelian Thesis is absurd if it does not exempt arguments that have undetachable or conditional conclusions. It does not follow from (XI) or (XIV) that I should do something other than chatting, or that I should pursue beautiful things, or that I should do anything whatsoever. Rather the conclusions of these arguments are conditional statements, explicitly in the former, and implicitly in the latter. That is, the latter argument at most concludes that whenever nothing cancels this reason, and there are no overriding reasons not to pursue a beautiful thing, I should pursue beautiful things. The Aristotelian Thesis can be, however, only a thesis about *detachable* or *unconditional* conclusions. This restriction should not surprise us. The "job" of practical reasoning cannot end at a conditional conclusion. Insofar as one has not yet settled on a course of action, practical reasoning has not yet come to a rest, and thus one cannot see any such conclusions as any more than inferential steps in a larger piece of reasoning. Thus, more generally, we can say that the Aristotelian Thesis is a thesis about proper *termini* of practical reasoning, not about the conclusion of any thinking that has possible human ends as a subject matter. To restrict the Aristotelian Thesis in this way is just to clarify that the relevant notion of a "conclusion" here is the notion of something that can be regarded as a real *terminus* of reasoning that is indeed *practical* (as opposed to something that could be the end point of idle speculation).

Let us go back to our first argument with this clarification in mind, with the conclusion explicitly stated as an intention as in (VIII). Is (VIII) an unconditional conclusion? It seems that if this is a valid piece of reasoning, the answer is "no." Let us assume again that the premises are all sound. For suppose I were also aware that the children are asleep and I will wake them up if I turn the light on, and that I would like to read, but I much prefer not to wake the children up than to read. In this case, the conclusion of the argument would specify an intention to act against my preferences. In accordance with (C_1), we cannot consider the argument as it stands to be valid and unconditional. A fully rational agent would not choose to act in way that obviously goes against his own preferences. As we said above, the validity of the inference depends on adding a *ceteris paribus* clause. So far we must read the conclusion as "If everything is equal, I intend to turn the light on" as a conclusion.[12] So, if we want to assess the truth of the Aristotelian Thesis, we should investigate an argument that does not have a *ceteris paribus* clause. Can we get rid of the *ceteris paribus* clause, while still holding on to the view that the conclusion of the reasoning is an intention? One simple way to do this is to add to the reasoning the following premise:

IIa. Everything else *is* equal.

Adding this premise seems, at first, to do the job. One can complain that this does not present the agent's full reasoning, since it does not say why the agent thought that everything else *was* equal. But even this flaw can be perhaps fixed if we move away from presenting the agent's reasoning as a form of practical syllogism. Indeed the practical syllogism seems to capture only a fraction of the agent's reasoning. An agent typically weighs various pros and cons of a situation, and, one could argue, a proper representation of practical reasoning should bring to light this kind of procedure. Moreover we could get rid of the *ceteris paribus* clause altogether by registering all the relevant considerations. Now, this seems hardly feasible in practice,[13] but at least it can show us how a valid piece of practical reasoning that would have an unconditional conclusion.

---

[12] Another way to secure the validity of the argument with an unconditional conclusion is to argue that practical reasoning is non-monotonic. See on this issue, Brandom (2001). I leave this possibility aside for the moment and come back to it at the end of the paper.

[13] Especially if we think that part of the reasoning involved registering indifference about various things. After all it is not a matter of logic, for instance, that moving my left foot first when I start walking towards the switch is no better than moving my right foot first, and thus, arguably the full representation of the reasoning would involve listing every single aspect of the action the agent is about to undertake (or at least every single aspect that the agent does or could foresee) in comparison to the alternatives.

If we now add (IIa) to the reasoning, or further premises specifying all the relevant considerations that weighed in the decision to turn the light on, can (VIII) be rightly considered the *terminus* of practical reasoning? Now even if one thinks that the conclusion of practical reasoning is an intention, one will not think that the agent's practical life ends at the formation of intentions; the point of forming intentions is to carry them out in actions.[14] Suppose now the straight path to my light switch goes through my computer, which I can easily, but damagingly to the computer, shove out of the way. Had I taken this route I would have carried out my intention, but my *action* would no longer be justified by the relevant piece of reasoning. For surely the reasoning left out the fact that it would not be worth trashing the computer. But if this is the case the reasoning does not warrant *this particular intention*. Assuming that I am fully aware that taking this route should knock the computer in this manner, I act against my preferences if I take this route, and thus irrationally. In accordance with (C_2), given that we need not assume that there is anything wrong in the way I form beliefs or desires, and since there is no unusual cognitive process that could take the blame for my irrationality, my irrationality must be due to bad reasoning. Indeed, it seems independently clear that the reasoning can make my intention rationally acceptable as it stands. For, unless our intentions have certain autonomous benefits,[15] our intentions can only be justified if we are justified in carrying them out. And in this case, I am not rationally warranted to carry out my intention, at least, not to carry it out in any possible way. Of course one could protest here that although not all ways of carrying out the intentions are justified by this piece of practical reasoning, at least some of them are. But this reply concedes that the intention is not the proper *terminus* of practical reasoning: insofar as we want to allow that some instances of practical reasoning are valid, the job of practical reasoning is not done when we form this intention, for, given that not just any way of carrying out the intention is warranted by the reasoning above, in order to know how to act I must also know which ways of carrying out the intention would be appropriate.

Still, the critic of the Aristotelian Thesis might argue that this simply shows that the intention needs to be further specified. First one might say that given that the agent was aware of the existence of the computer in

[14] For instance, Thomson claims that the Aristotelian Thesis is "at best suspect," but later argues that practical reasoning is reasoning "about what to *do*" (Thomson 2001, pp. 79 and 82, emphasis added).

[15] Such as the benefits of forming the intention to drink toxin in Gregory Kavka's toxin puzzle. See Kavka (1983).

the path to the light switch, this must have been a relevant consideration in forming the intention, and thus should be part of the content specified by the intention. But this will not do. First, note all that needs to go into the content of an intention if I need to represent all the foreseeable ways in which carrying out the intention might be acceptable or unacceptable. If I am going to turn on the light in my room, and now I am in the next room over (a rather simple intention to be executed), I must represent the layout of the room, and my path towards the room, make sure that I keep in mind all possible obstacles, represent how I will move my arms and legs so as to avoid the possible obstacles, think about what can happen in my room that can make turning the light on in a certain way problematic, represent more precisely *how* I am going to turn on the light, etc. It is quite implausible that this is all even implicitly represented in forming the intention to turn the light on; indeed implausible enough that we might want to reconsider the plausibility of the Aristotelian Thesis. But suppose one were ready to bite the bullet here, and argue that I do represent all these things when I form the intention to turn the light on. This still will not suffice. Insofar as the intention *precedes* the action,[16] and the action is extended through time, the agent could always become aware of new, relevant information *while* he executes the intention. It seems that any intention that can be the conclusion of practical reasoning must take into consideration the possibility that in executing an intention, the agent might face unexpected, but relevant, facts. One could try to handle these problems in one of two ways. One could first try to make the intention so specific that it will rule out the possibility of unexpected "twists" while one acts. In this case, the intention could be something like:

VIIIa. I intend to walk through such and such a path to reach the light switch and then turn the light on.

This is a short-lived improvement. For specifying the path is not all I need to do. I could take this path but flip the switch with my mouth, both experiencing the (I imagine) unappetizing flavour of light switches, and exposing myself to germs in such a way that may outweigh the value of switching the light on. The problem is that by just getting more and more detailed about my plan cannot rule out in advance that, while I act, I will

[16] Could one say that the intention that is the conclusion of practical reasoning does not *precede* the action, but is an intention *in* action? I discuss this point below, but in a nutshell this would surrender most of the advantages that the view is suppose to have over the Aristotelian Thesis, and would make it virtually indistinguishable from the Aristotelian Thesis.

face previously unforeseen information that is relevant for how I act. The general problem should be clear: the intention that is supposed to be the conclusion of practical reasoning, the intention that guides me in action, is the representation of something general. But the action is a particular. Thus there always are aspects of the action that were not represented in the intention, of which I become aware while I execute the intention. Since each aspect is potentially relevant for the evaluation of my action, the way I carry out my intention can never be fully justified by the practical reasoning that issues in this intention. It might be thus better to deal with the unexpected by representing it in a general form in the agent's intention. Perhaps the relevant intention is something like the following:

VIIIb. I intend to turn the light on while always making sure that, as far as I can tell, no foreseeable effect of my carrying out the intention in a particular way outweighs the value of the action.

(VIIIb) succeeds in covering the whole ground by quantifying over all foreseeable effects of my action. However (VIIIb) is also a conditional conclusion, or at least a conclusion that leaves the job of practical reasoning unfinished. It has a form equivalent to "I intend *X* unless *C*," or "I intend to *X* in some way" (but the correct way of doing it still needs to be figured out), and for this reason it cannot determine in any particular way how I should act. In sum, no matter how one further specifies the intention, given that the intention is general and the action is particular one will be facing the following dilemma. On the first horn, one would say that *any* way of carrying out the intention specified in the conclusion would be justified by a piece of sound practical reasoning. But this route is hopeless; given that there are indefinitely many ways of carrying out an intention one would expect that some of them could turn out not to be rationally justified. On the second horn, one would say that *only some particular ways* of carrying out the intention are justified by a piece of sound practical reasoning. But if this is the case, the intention can't be the terminus of practical reasoning, for one cannot yet act in a justified manner until practical reason can specify *which* particular ways of carrying out an intention are justified.

One could say that the problem here is not with further specifying the intention, but with individuating the appropriate stages of one's behaviour such that a specifiable intention corresponds to each. I have assumed that while carrying out an intention one might become aware of certain things one hadn't been aware of (or couldn't have foreseen) at the time that the intention was formed, but that, once one is made of aware of

them, this would render certain ways of carrying out the intention irrational. If I could not foresee an obstacle when I formed my intention to turn on the light (for instance, there are shards of glass in my way to the light that cannot be spotted from where I was), but have become aware of it while executing the intention in time to avoid to obstacle, it would be irrational of me to continue to carry out the original intention in such a way as not to avoid the glass shards. However, one may object that this assumption is plausible only if we do not ascribe a separate intention to each "choice node." We do not turn the light on by merely directing our will towards this end, but we take steps in the direction of the light switch, we move our hand in the direction of the switch, we press it down, etc. Each of these steps presents a choice situation, in which we need to make a decision; each of these steps, the objection goes, requires a separate intention. Since there are no choice nodes between each of these intentions, there is nothing that I can become aware of between the time I form each of these intentions and the time I finish executing them that could make a difference to the rationality of my actions, since there is nothing I can do between choice nodes.

Now, I do think that this is the most promising way to reply to the objection. No doubt the number of intentions that need to be postulated will be quite high, but given that part of the problem here is that there seems to be so much reasoning that goes into an action, this crowding of intentions might not be so objectionable. It would also be unfair to protest that it is it hard to believe that we think to ourselves each of these intentions after explicit deliberating about it in our mind. After all, it is hardly a minority view that some intentions and some deliberations do not show up in the agent's life as explicit, occurrent thoughts.

However, this option will end up facing a few serious problems. The first one is that it is not so clear how different this proposal is from the Aristotelian thesis itself. It is hardly likely that this approach would succeed in postulating a *separable* intention for each stage of the action. Rather it would probably rely on the existence of what is sometimes called "intention in action"; that is, an intention that is an aspect of my intentional action, not an event that can be separated from it. Given the omnipresence of such "stages" at almost any moment in which one is carrying out a continuous action, it seems that in this approach we need to look at the decision embodied *in* carrying out the action at each step, rather than the decision *to* carry out the action. But in this case, the gains of moving to this conception of practical reasoning are rather limited; the conclusion of practical reasoning here is still inseparable from the action itself. And if the conclusion of practical reasoning is something that is

inseparable from the action itself – indeed something that can be described as an aspect of the action – one seems to have gotten quite close to conceding, if not fully conceded, that the conclusion of practical reason *is* an action. Indeed, the main advantages of taking intentions to be the conclusions of practical reasoning, at least in dealing with objections (2) and (3), seemed to rest precisely on the fact that it took the conclusion of practical reasoning to be separable from the action, something that could occur even when reasoning did not issue in an action. Moreover this view would have to find a way to parse the relevant intentions in continuous actions. For instance, when I am running, it seems that I can decide to stop at any moment, and at any moment my failure to do so could be a failure of deliberation.

Indeed this problem becomes particularly difficult when we try to understand how this suggestion would deal with the skilful execution of an intention. Let us look at two tennis players, a rather skilled one, and one who is learning to make the shots. For our present purposes, the second player is going to be an idealization, since we will assume that he proceeds by explicitly reasoning how to turn his hand, how to place his racket given the trajectory of the ball, etc., and still has time to make the shot. They both enter the court with the intention to win the game, and they will form, on this view, various more specific intentions throughout the game. Suppose now they face the exact same situation: the opponent returns a serve in such a way that she leaves one side of the court completely open. Both come to the same conclusion about what to do in this situation: each must send a hard shot to that side of the court. Now the unskilled tennis player cannot just hit the shot; the job of practical reasoning is not yet over for him. He must try to figure out the approximate speed and trajectory of the ball, calculate the angle he wants his racket to be at, how hard he has to hit it, etc. For the unskilled player, this view is no doubt committed to saying that these were stages of the action that required further deliberation and further intentions. But what about the skilled player? Here this view will face a dilemma. On the one hand it seems that we are committed to saying that there are no further stages of this action. For a certain period of time, the skilled player will not have acted under any intention other than "hitting a hard shot on the deuce side." If asked why he had his racket facing down at a sharp angle, he might not even recognize that he did anything that fell under this description. Indeed, if the skilled player misses the shot, it would be a bad shot. It would not be an instance of irrationality; it would not be an instance of failing to deliberate correctly about how to carry out the intention to hit a hard shot on the deuce side. If we look back at (C_2),

none of conditions (a)-(d) seem to apply to this case. The same is not true, or at least not necessarily true, of the unskilled player. The unskilled player settles on the position of the racket by deliberating on the issue, and thus, at least in the case in which he is aware of all the relevant information, if the unskilled player chooses the wrong angle, she will have deliberated badly. Thus since there were no more specific intentions about how to carry out the intention to hit a hard shot on the deuce side, we seem to come to a conclusion that there are no further stages of the action. However, this view seems also equally committed to the existence of further stages of this action. For the agent could change the course of the action at any moment, and this fact could be relevant for evaluating whether the conclusion of her practical reasoning was warranted or acceptable. One could, for instance, ask: "Didn't you notice that there are children running across the court all the time? And that a child whom you couldn't see at the time you decided to hit the shot could run to the deuce side of the court and be there just in time to be hit by the ball? Do you care more about winning a game than about the welfare of a child?" It is certainly possible that the appropriate answer here is something like: "Any such child would have to appear in my field of vision before my racket hit the ball, in which case I would just send the shot in a different direction." No doubt the availability of such an answer is relevant to the acceptability of the agent's reasoning, and thus it seems that we need some reasoning that has the conclusion "I can go ahead and hit my shot" just before the skilled tennis player hits the shot. But since the conclusion of practical reasoning under this view is always in an intention, so it seems that we are at same time, under this view, required to say that there *are* further stages in the action of the skilled player. The problem is that skilful execution of an intention is a way in which we carry out an intention such that we are still in control of our actions (and we still could thus revise the intention) but not by means of further, more specific intentions, as would be required by the view in question; in these cases, one carries out an intention without representing the way in which one is carrying out the intention.[17]

[17] No doubt one could continue the argument here, by trying to say, for instance, that the further stages can be characterized as various intentions to continue carrying out the original intention. I do not think that this strategy would work since we would need more determinate ways of specifying the intention to continue, and given the nature of skilful execution of intentions, this might not be possible. But at any rate, my aim is to establish that such a view would have enough problems that it is worth reconsidering the Aristotelian Thesis.

One might be tempted to do away with these problems by advocating a hybrid view: that the conclusion of practical reasoning is sometimes an intention and sometimes an action. But it seems hard to prevent this concession from turning into full surrender. For after all most of our actions are stretched through time, and most of them require some kind of skilful execution of one's intentions. Before we set ourselves to protect such an enclave for the view that intentions are the conclusion of practical reason, we should re-examine the plausibility of the Aristotelian thesis in light of our revised understanding of its subject matter. If the Aristotelian Thesis can answer these questions, the issue of whether or not such an enclave can be protected might be moot.

4. The Objections against the Aristotelian Thesis Reconsidered

If the above arguments are sound, a piece of practical reasoning will not be able to justify an unconditional intention, since particular ways of carrying out the intention (and in some cases all particular ways of carrying out the intention) will turn out not to be warranted by the apparently valid piece of practical reasoning. In general, we see that what needs justification is not only the general end represented in the intention, but the particular way in which one carries out the intention; indeed, practical reasoning ideally should justify that no particular way of carrying out the action would be more advisable. Thus the only thing that could be properly warranted as the unconditional conclusion of practical reasoning is the particular way of carrying out an intention, and thus the action itself. What the practical syllogism justifies is the particular action carried out by the agent, not the intention itself. (I) and (II) justify my *particular action* of turning on the light, but they could not justify the intention, since many ways of carrying out the intention would not be warranted in light of such beliefs and desires.[18] The more perspicuous way of writing (III) would be:

IIIa. This particular action of turning the light on.

There is no issue here of a *ceteris paribus* clause; since the conclusion is the action itself, either it is justified, and thus there was nothing that made it unwarranted, or it is not justified, and thus the inference is invalid. This approach also provides us with a quite straightforward response to the objections raised against the Aristotelian Thesis. First,

[18] Although, again, a conditional intention could be justified.

although it is true that we intend the action only under a particular description, whether the action is *justified* must take into account more than the description under which I intended it. For at least unintended but foreseen consequences will play *some* role in assessing the soundness of my reasoning.[19] Moreover, assuming that (I)-(IIIa) is valid, and that by turning the light on I alerted the prowler, rewriting the conclusion as follows would still give us a valid inference:

IIIb. This particular action of alerting the prowler.

Since the action itself was justified by (I)-(II), picking out by means of a different phrase could make no difference to this fact; the fact that (IIIb) can also pick out the conclusion of my action, for instance, certainly doesn't violate (C_1); a fully rational agent could perform the action describe in (IIIb), even she would not intend it under this particular description. No doubt writing (IIIa) as the conclusion makes the validity of the inference more perspicuous. This is all no different from theoretical reason: substituting equivalent propositions might turn an obviously valid argument into one whose validity only a skilled logician could establish.

Certainly we may fail to carry out an intention for reasons that have nothing to do with our reasoning capacities – for instance if I were to die before I could carry out my intention. This case is unproblematic for our account: were I to die I would never have derived the unconditional conclusion.

Moreover we can understand the case in which I clumsily do something other than what I intended as a case in which the conclusion of practical reasoning is indeed an attempt such as:

VIIa. This particular action of trying to turn the light on.

Since there is no doubt here that 'trying' here refers to the "outward" action, there is no danger of having the conclusion slide back into a mental state.

One might also complain that the reasoning from (I)-(IIIa) cannot represent the full reasoning of the agent. For, after all, (I)-(II) could not justify the actions by themselves. Had it been the case that I knew I would have electrocuted myself by flipping the light switch, I would not be justified in turning the light on, even if I wanted to read. But although this is true, I do not think it follows that *in this case* (I)-(II) were not

[19] This is not to say that it makes no difference whether an effect is intended or foreseen. However, even the most adamant defender of the doctrine of double effect will not argue that we should simply disregard foreseen but unintended consequences.

sufficient to justify the conclusion. Since the conclusion is a particular action, and not a general claim about what one ought to do in these circumstances, the truth of that counterfactual does not affect the validity of the inference. Since the inference is not supposed to justify a certain general *description* of an action, the fact that another action falling under the same description would not be justified by the same premises is irrelevant to assessing the acceptability of the inference in question.[20] Could not the same be said about the account that takes an intention to be the conclusion of practical reasoning; that is, that the inference warrants the intention to turn on the light on this particular occasion? But here again the intention that I can form even on this particular occasion is still a conditional one: the intention spelled out at (VIIIb). There is no escape from the fact that practical reasoning comes to a rest only when the action is completed. Thus anything that stops short of the action itself must be the intervening chapters, rather than the conclusion, of practical reason.

One might be surprised here at the disanalogy between theoretical and practical reasoning. After all, it is also true that if it rains, it rains in a determinate way. But this neither seems to affect our views about the acceptability of conclusions such as "It will rain tomorrow" nor does it make us think that these are not proper resting points for theoretical reasoning. However, we can see why this disanalogy holds. The aim of theoretical reasoning is knowledge or true belief. But if it is true, say, that it will rain heavily tomorrow, it is still true that it will rain tomorrow. The fact that it will rain tomorrow in a particular way does not make the statement "it will rain tomorrow" any less true. But the aim of practical reasoning is "right" or "justified" action. But it does not follow, for instance, from the fact that one is justified in eating that one is also justified in eating heavily.[21]

Acknowledgements

This paper was presented at the University of Bristol, at Universidade Federal do Rio de Janeiro, at 2003 Central Division of the American

[20] This does show that inference is at best a materially, rather than formally, valid inference, and that the inference is non-monotonic. On a similar point, see Brandom (2001).

[21] Perhaps there is a closer analogy here with being justified in asserting the conclusion of a piece of probabilistic reasoning. Pursuing the analogy, however, would require an investigation of the nature of probabilistic reasoning going beyond the scope of this paper.

Philosophical Association, and at the 2002 meeting of the Canadian Philosophical Association. I would like to thank the audience in all these occasions for valuable comments and especially Nancy Snow, my commentator at APA meeting, and Richmond Campbell, my commentator at the CPA meeting. I also would like to thank Donald Ainslie, Phil Clark, Jimmy Doyle, Jennifer Nagel, and Fred Schueler for their insightful comments on earlier drafts of the paper.

University of Toronto
Department of Philosophy
215 Huron St.
Toronto, M5S 1A1, Canada
e-mail: sergio.tenenbaum@utoronto.ca

REFERENCES

Aristotle (1984). The Movement of Animals. In: *The Complete Works of Aristotle*, pp. 1087-1096. Edited by J. Barnes. Princeton, NJ: Princeton University Press.

Audi, R. (1989). *Practical Reasoning*. Oxford: Oxford University Press.

Brandom, R. (2001). Action, Norms, and Practical Reasoning. In: E. Millgram (ed.), *Varieties of Practical Reasoning*, pp. 465-479. Cambridge, MA: The MIT Press.

Clark, P. (1997). Practical Steps and Reasons for Actions. *Canadian Journal of Philosophy* **27**, 17-45.

Davidson, D. (1980). Actions, Reasons, and Causes. In: *Essays on Actions and Events*, pp. 3-20. Oxford: Oxford University Press.

Feinberg, J. (1965). Action and Responsibility. In: M. Black (ed.), *Philosophy in America*, pp. 29-45. Ithaca, NY: Cornell University Press.

Harman, G. (1986). *Change in View*. Cambridge, MA: The MIT Press.

Hornsby, J. (1998). *Simple-Mindedness*. Cambridge, MA: Harvard University Press.

Kavka, G. (1983). The Toxin Puzzle. *Analysis* **43**, 33-36.

Korsgaard, C.M. (1997). The Normativity of Practical Reasoning. In: G. Cullity and B. Gaut (eds.), *Ethics and Practical Reason*, pp. 215-254. New York: Oxford University Press.

Mele, A. (1992). *Springs of Action*. New York: Oxford University Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Thomson, J.J. (2001). *Goodness and Advice*. Princeton, NJ: Princeton University Press.