OXFORD

oxford studies in normative ethics

Oxford Studies in Normative Ethics

Oxford Studies in Normative Ethics

Volume 6

EDITED BY
MARK TIMMONS

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2016

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2011294439

ISBN 978–0–19–879058–7 (hbk.)
ISBN 978–0–19–879059–4 (pbk.)

Printed in Great Britain by
Clays Ltd, St Ives plc

Contents

Acknowledgments vii
List of Contributors ix

Introduction 1
Mark Timmons

1. Taking Account of Character and Being an Accountable Person 12
Stephen Darwall
2. Taking Pride in Being Bad 37
Claudia Card
3. Character as a Mode of Evaluation 56
Kate Abramson
4. The Normative Force of Promising 77
Jack Woods
5. Promissory Obligation: Against a Unified Account 102
Hallie Liberto
6. Two Concepts of Rule Utilitarianism 123
Susan Wolf
7. After Solipsism 145
David Schmidtz
8. Extrinsic Value and the Separability of Reasons 166
Barry Maguire
9. The Relativity of Ethical Explanation 189
Kenneth Walden
10. Two Senses of Moral Verdict and Moral Overridingness 215
Paul Hurley

11. Love in Spite of 241
Erich Hatala Matthes

12. Moral Reasoning 263
Gilbert Harman

Index 277

Acknowledgments

Versions of eleven of the twelve chapters in this collection were presented at the sixth annual Arizona Workshop in Normative Ethics that took place in Tucson, Arizona on January 16–18, 2015. Claudia Card was to have been a keynote speaker at the workshop, but became ill and was not able to participate. She died on September 12, 2015. The paper she was to present, "Taking Pride in Being Bad," is included in this volume. I would like to thank the co-executors of Professor Card's literary materials, Victoria Davion, Kathryn J. Norlock, and Lynne Tirrell, for making it possible to include the paper.

I thank the Center for the Philosophy of Freedom, the Department of Philosophy at the University of Arizona, and the John Templeton Foundation for their generous financial support of the workshop. Of course, the views expressed in the volume's chapters do not necessarily reflect the views of the Center, the Department, or the John Templeton Foundation.

Thanks to Chris Howard who assisted me in running the workshop, to Lucy Schwarz for preparing the volume's index, and to Betsy Timmons for her gracious help with workshop details. I would also like to express my sincere thanks to the following philosophers for serving as de facto program referees: Gwen Bradford, Jennifer Hawkins, Robert Johnson, Doug Portmore, Luke Robinson, and Eric Wiland. Two anonymous referees for Oxford University Press offered very helpful, constructive advice to our authors. Thanks finally (and again) to Peter Momtchiloff, my OUP editor, for his support.

Mark Timmons
Tucson, AZ

List of Contributors

Kate Abramson is Associate Professor of Philosophy at Indiana University

Claudia Card was the Emma Goldman Professor of Philosophy at the University of Wisconsin–Madison

Stephen Darwall is Andrew Downey Orrick Professor of Philosophy at Yale University

Gilbert Harman is James S. McDonnell Distinguished University Professor of Philosophy at Princeton University

Paul Hurley is Edward J. Sexton Professor of Philosophy at Claremont McKenna College

Hallie Liberto is Assistant Professor of Philosophy at University of Connecticut

Barry Maguire is Research Assistant Professor at University of North Carolina, Chapel Hill

Erich Hatala Matthes is Assistant Professor of Philosophy at Wellesley College

David Schmidtz is Kendrick Professor of Philosophy at the University of Arizona

Kenneth Walden is Assistant Professor of Philosophy at Dartmouth College

Susan Wolf is Edna J. Koury Distinguished Professor at University of North Carolina, Chapel Hill

Jack Woods is Assistant Professor of Philosophy at Bilkent University

Introduction

MARK TIMMONS

Oxford Studies in Normative Ethics aims to publish cutting edge work on a range of topics in the field of normative ethical theory. This sixth volume brings together twelve new essays that collectively cover a range of fundamental topics in this field, including: the nature and morality of character, moral evil, the wrongness of promising, forms of utilitarianism, value and reasons, ethical explanation, moral overridingness, love, and moral reasoning.

It is common in philosophical ethics to distinguish aretaic assessment, whose focus is on motives and traits as they express one's character, from deontic assessment whose focus is an agent's actions. As **Stephen Darwall** explains in "Taking Account of Character and Being an Accountable Person," these two forms of assessment correspond respectively to responsibility as attributability and responsibility as accountability. Responsibility as accountability invokes such Strawsonian reactive attitudes as guilt, resentment, indignation, and blame, and Darwall argues that there is a tight conceptual connection between accountability concepts and deontic concepts. Specifically, the tight connection here is mediated by the concept of blameworthiness. Moreover, we are accountable to others as "representative moral agents, where moral agency is understood to include the capacities necessary to enter into relations of mutual accountability." Because deontic assessment has this connection with blameworthiness and being accountable to others, it is essentially second-personal. In contrast, responsibility as attributability, concerned with assessment of character and employing such evaluative concepts as esteem and disesteem, is essentially third-personal. Given the difference between these two notions of responsibility, one might suppose that accountability does not apply to motives, dispositions, and character generally. But this is not Darwall's position. Rather, he argues that having certain dispositions, such as being excessively self-centered or

viewing oneself as somehow specially privileged in relation to others, is likely to interfere not only with one's compliance with moral obligations, but also likely to distort one's capacity for moral perception and judgment. For these reasons, argues Darwall, one may be held accountable for taking account of one's character. Thus, while the notion of accountability applies in the first instance to acts apart from character, it extends to character.

In "Taking Pride in Being Bad," **Claudia Card** attempts to make sense of the idea of valuing something in virtue of its badness, indeed, of taking pride in being a "badass," in being someone who takes pride in having "those qualities of character that enable one to be good at being cruel, hard-hearted, merciless, ruthless, manipulative, making people suffer, terrifying people, and so forth." The sort of badass Card has in mind is someone who takes pride in doing evil not simply as a means of gaining something else one values, such as the recognition and esteem of select others. Rather, the sort of evil character in question is someone who aims to be *worthy* of the approval of other badasses, by valuing cruelty, ruthlessness, and other such evils for their own sake. Famously, Kant denied that human beings were capable of evil for evil's sake, of having a "diabolical will" of the sort that would take pride in being bad. Card rejects Kant's view of evil and proposes as an alternative an "atrocity paradigm" of evil that makes room for taking pride in being bad. In explaining how a person can come to have such an evil character, Card appeals to Christine Korsgaard's Kantian conception of a practical identity and to "attachment theory" in psychology. Korsgaard's Kantian conception allows for a wider range of individual self-conceptions than Kant's view allows, and thereby makes room for having a self-conception of the sort we find in the badass. Attachment theory, according to Card, can be used to amplify Korsgaard's view by explaining how one's self-concept is greatly influenced by one's early interactions with others who have been (and perhaps continue to be) important in one's life. According to psychologist Lorna Smith Benjamin, this sort of attachment to others can explain perverse, irrational, or downright diabolical behavior. Thus, one explanation of the badass, and one being proposed by Card, is that such a person, via attachment to someone else who is perceived as a badass, comes to emulate that person and thereby comes to take pride in being bad. Card concludes by noting that Kant's view of evil was partly correct in supposing that a predisposition to evil is not an innate element

of human nature; that coming to form an attachment to an evil model is not something *initially* diabolical. However, because Kant did not recognize a predisposition to form attachments to others, his moral psychology could not explain how it could be that someone ever comes to take pride in being bad.

Character traits, including virtues and vices, are standardly treated as distinct *kinds* of psychological attribute, distinct from other psychological attributes such as forms of mental health and illness as well as natural abilities and inabilities. In her "Character as a Mode of Evaluation," **Kate Abramson** challenges the standard view. She argues that conceiving of character traits, natural abilities/inabilities, and aspects of mental health and illness as being distinct psychological kinds, results in a taxonomy that fails to correspond to our shared practices of psychological classification. Abramson proposes that the core differences at issue concern *modes of evaluation*, rather than psychological kinds. She argues that the "question of whether we should regard any given psychological attribute as an aspect of character, an aspect of mental health or illness, as a natural ability or defect, as a talent or skill is . . . a choice, at root, amongst modes of evaluation." Such modes of evaluation—moral mode, medical mode, and natural ability mode—differ in content, implications, and appropriate conditions of application. For instance, distinctive of the moral mode is that persons properly evaluated according to it are fitting subjects of reactive attitudes such as praise and blame. Further, in some cases, more than one mode of evaluation is properly applicable. Someone who is a compulsive liar is appropriately evaluated by both moral and medical modes. In developing her modes of evaluation proposal, Abramson discusses the appropriateness conditions for the various modes of evaluation and how those conditions explain why it is sometimes fitting to evaluate a single attribute under more than one mode, but sometimes not. Her proposal is that choice of modes in a particular context will likely depend on a complex interplay among the following factors: (1) psychological facts about an agent and the fit between them and one or more of the modes, (2) the interpersonal import of adopting some mode, and (3) the comparative appropriateness of adopting one mode rather than another in light of the first two sorts of factor.

What explains the normativity of promising—that from the fact that one has promised to do something, one thereby has a reason to do it?

This is the question **Jack Woods** attempts to answer in "The Normative Force of Promising." The most seemingly plausible answers to this question, according to Woods, are conventionalist in the sense that they make essential reference to the value of the practice in explaining the normativity of promising. David Hume, T. M. Scanlon, David Owens, and Brad Hooker are all proponents of the conventionalist explanation (though, of course, the details they offer in their explanations differ). However, such views are subject to counterexamples (e.g., death-bed promises) and so cannot provide a complete explanation of the normativity of promising; they lack explanatory scope. Moreover, some of these views cannot accommodate the so-called particularity of promissory reasons, that is, they do not make adequate sense of the fact that when A breaks a promise to B, B has a particular reason to object to the promisor's breaking the promise to *him*—a reason distinct from some general reason anyone might have for wanting promisors to keep their promises. To overcome these difficulties with standard conventionalist views, Woods proposes what he calls a "quasi-conventionalist" account of the normativity of promising. One key component of this view is the idea of blame-liability, a feature that is present in those cases that are problematic for standard versions of conventionalism. However, according to Woods, this feature of conventions governing promising is not enough to explain the normativity in question. What Woods proposes in securing the explanation is a desire-based account of reasons, according to which one has reasons to satisfy one's desires. Given that one has a sufficiently strong reason to avoid being blame-liable, and that keeping one's promises serves to avoid being blame-liable for violating the conventions of promising, one thereby has an instrumental reason to keep one's promises. This is quasi-conventionalism because although the conventions governing promising play an essential role in the desired explanation, the normativity of promising is grounded in one's desire-based reasons. If, as Woods claims, his quasi-conventionalism accommodates both the exceptional cases that confound standard conventionalist views as well as the particularity of being blame-liable for breaking promises, then it emerges as superior to competing views. Having set forth his quasi-conventionalism, Woods concludes by replying to various objections.

However, whether there is a single account of the normativity of promising is the topic of **Hallie Liberto's** "Promissory Obligation:

Against a Unified Account." Unified accounts, including the Expectation model, the Reliance model, the Authority model, and the Trust model, all face challenging counterexamples. Of course, from the presumptive fact that each of these accounts has counterexamples, Liberto's non-unification thesis does not follow. So, after critically discussing the leading unifying accounts, she goes on to give a positive argument for her thesis. Her general strategy is to consider a pair of promises that have the same spoken content, are made in the same context, and made with the same intentions, and then argue that the promissory obligations in the two cases are not grounded in a single explanatory feature. Abstractly described, for one of member of the pair, it is arguably the role of *reliance* that explains the particular range of exclusionary conditions (conditions under which one is released from the promise), and also explains what constitutes breaking the promise. And so in such a case, it is plausible that the duty not to forsake an invited reliance is what grounds the promissory obligation. However, in the paired example it is not the duty to forsake an invited reliance that explains the exclusionary conditions or what counts as breaking the promise, and so it is not the invited reliance that grounds the promissory obligation in the case at hand. The moral that Liberto draws from this type of example is that there is more than one type of promissory obligation and so one should not expect there to be a completely unified account of such obligation.

The alleged advantage of rule (restricted) utilitarianism over act (extreme) utilitarianism is that it saves utilitarianism from implausible implications about particular cases, such as scapegoating an innocent person in cases where doing so would maximize utility. But as J. J. C. Smart pointed out many years ago, complying with justified rules in cases where one knows that complying with them would fail to maximize utility amounts to mere superstitious rule-worship. In "Two Concepts of Rule Utilitarianism," **Susan Wolf** explores how the objection might be met by distinguishing two conceptions of morality. The *moral point of view* conception specifies those considerations that count as reasons as well as the relative weights of those reasons, and Wolf maintains that if one works with this conception and embraces utilitarianism (which combined she calls the standard conception), then J. J. C. Smart's objection to standard rule utilitarianism is seemingly unanswerable; from the utilitarian perspective wedded to morality as a point of view,

individual actions as well as rules are to be evaluated strictly in terms of their effects on the common good.

The way around the rule-worship objection, according to Wolf, is to replace the point of view conception of morality with what she calls "the practice conception," inspired by Rawls' "Two Concepts of Rules." According to this conception, morality is a "loose and informal" set of practices constituted by rules that specify offices, roles, penalties, as well as duties and obligations. Returning to Smart's rule-worship objection, Wolf proposes that by embracing the practice conception a rule utilitarian can answer Smart by appealing to such non-moral, personal justifications as: repugnance at the idea of claiming privileges in relation to breaking moral rules, the satisfaction of living with others on equal and open terms, and living up to one's cherished ideals. Importantly, on the practice conception, these reasons in ordinary contexts of decision-making are not weighed impartially along with considerations of utility as they would be from the morality as a point of view perspective. And this allows the rule utilitarian to answer Smart's rule-worship objection by holding that (1) the purpose of moral rules and morality generally is to bring about the greatest good, (2) yet from one's own perspective this is not one's own purpose in living. If Wolf is right, then the practice conception of rule utilitarianism can both avoid the problematic cases that act utilitarianism faces and yet avoid the charge of rule-worship. Wolf concludes with the historical speculation that Mill's defense of utilitarianism, which has struck many interpreters as ambivalent between act and rule versions, is best understood as implicitly committed to the practice conception of morality.

In "After Solipsism," **David Schmidtz** asks how best to conceive moral theory given that we live in a strategic world—a world in which we have ongoing associations with other individuals, where one's own choices affect and are affected by the choices of others. A moral theory, one of whose aims is guidance in fostering human flourishing, needs to take seriously the need for productive cooperative ventures. According to Schmidtz, this means taking seriously the need for moral theory to reject a parametric model of practical decision-making that ignores realities of our strategic world by inviting one to reason as if one were acting alone to produce good outcomes. One such moral theory, based on a parametric model, is a particular interpretation of act utilitarianism that requires individuals living in a world with poverty to aim at maximizing

overall utility by reducing oneself to the level of marginal utility by a policy of "unconditional giving." Critics typically charge this kind of theory with being too demanding. In sharp contrast, Schmidtz finds its model of decision-making too undemanding—undemanding of others—because it does not take seriously enough the idea that in our social world in which reciprocating is an ideal, the problem is one of "specifying terms of engagement that make separate persons willing and able to trust each other enough to launch and sustain society as a cooperative venture." Indeed, as Schmidtz points out, asking individuals to act alone in an effort to do what they can do to maximize utility threatens to encourage others to free-ride. The way to overcome this kind of solipsism in the practical realm, and fruitfully pursue the ideal of cooperation, is by developing social institutions that fully appreciate our strategic world and are sensitive to human history regarding what works in fostering human flourishing. As Schmidtz sees it, institutions (and practices generally) that are wealth-creating express a sense of cooperative venture for mutual advantage that makes it advantageous for individuals to act in ways that foster the common good. For consequentialists, then, moral life in a strategic world requires that one embrace *strategic consequentialism*. Schmidtz concludes by drawing the same lesson for Kantian deontologists. Living in a strategic world calls upon the deontologist to identify maxims that are "fit for a kingdom of players." Doing so would be to replace act deontology with *strategic deontology*.

Barry Maguire, in "Extrinsic Value and the Separability of Reasons" addresses a puzzle for act consequentialist theories that combine a value-based account of the deontic realm with a particular value-based account of virtue. For the consequentialist, from among one's options in a particular circumstance, one is morally required to perform actions that would maximize final value in that circumstance. According to a value-based conception of virtue, one's positive attitude toward some state of affairs is of final positive value if that state of affairs is itself finally valuable, and one's negative attitude toward some state of affairs is finally disvaluable if the state of affairs is of final positive value. But as Maguire points out, for consequentialist theories that embrace this particular value-based explanation of virtue, there can be cases in which, for example, committing a wrongful murder can turn out to be obligatory so long as the final value generated by enough individuals responding appropriately (and thus virtuously) toward the murder sufficiently

outweighs the disvalue of the murder. According to Maguire, then, the tension between value-based accounts of the deontic and particular value-based accounts of virtue arises in cases that feature non-instrumental extrinsic value—the sort of value realized, for example, in sadness as a response to tragedy. Maguire's central aim in this chapter is to develop a conception of the relation between value, reasons, and deontic status that avoids unwanted intuitive results concerning non-instrumental extrinsic value that arise for act consequentialism, while preserving a central motivation of consequentialism, namely, a value-based conception of reasons for action.

In "The Relativity of Ethical Explanation," **Kenneth Walden** defends the claim that ethical explanations are essentially contrastive in the sense that adequate explanations of ethical facts are given against a "space of foils" with which they contrast. To use Walden's own example, the adequacy of offering an explanation to one's interlocutor of why it would be wrong for someone to thrash his valet with a blackjack for some minor mishap will depend on, and thus be relative to, the contrast space being invoked. If one's interlocutor is interested in why it would be wrong for someone to thrash his *valet* as opposed to thrashing someone else, one contrast space is invoked, involving perhaps a butler or a cook. If one's interlocutor is interested in why it would be wrong to do the thrashing *with a blackjack* instead of some other instrument, a different contrast space is invoked, and so on for other interests. Part of Walden's defense of the contrastive nature of ethical explanation appeals to the contrastive, relative nature of explanation generally, including scientific explanation. Given this sort of relativity of ethical explanation, Walden then proceeds to draw two implications for doing ethical theory. First, the relativity in question brings into focus the possibility of ethically evaluating dubious explanatory contrasts that are embedded in ethical explanations grounded in one or another ethical theory. A second implication is that it can sometimes be a mistake in ethical theorizing to suppose that what seem to be competing ethical explanations of some phenomenon each aspire to provide a *complete* explanation of the phenomenon. That is, it may be that various distinct proposals to explain, for example, the appropriateness of partiality in certain cases need not be interpreted as representing *the* explanation; rather they should perhaps be evaluated for their adequacy against relevant contrast spaces, recognizing

that various "competing" accounts provide perfectly good explanations against one or another such space.

In "Two Senses of Moral Verdict and Moral Overridingness," **Paul Hurley** contrasts two senses of decisive moral verdict associated with two conceptions of moral overridingness. According to one sense, such verdicts reflect reasons for acting from a distinctively *moral standpoint*. According to another sense, decisive moral verdicts reflect decisive reasons that are distinctively moral reasons—the *rational standpoint* sense of moral verdict. One major point of contrast between the two senses is that according to the moral standpoint sense, it is a substantive question whether what is indeed required from the moral point of view is also rationally required. And so, according to this sense, moral requirements (or the reasons they reflect) are not necessarily overriding. By contrast, for those deploying the rational standpoint sense, whether an action is required from the standpoint of morality *just is* the question of whether the action is rationally required for reasons that are distinctively moral. According to the rational standpoint sense, then, moral requirements (or the distinctive moral reasons they reflect) are necessarily always overriding, though it is a substantive question whether there are such requirements. Hurley's chapter explores these contrasting senses of moral verdict, cautioning that to avoid a distorted understanding of the debate over moral overridingness, it is important to bring both senses into clear view and conduct moral theorizing in light of the distinction between them.

In loving someone, one must live up to certain requirements that are normative for love. In particular, one's love for another ought to be an attitude (or set of attitudes) one has in virtue of identity-forming properties which, from the perspective of the beloved, constitute her practical identity. One's love should be a response to who the beloved *is*—the identity requirement. But also, loving someone requires that one be especially concerned for the welfare of the beloved—the well-being requirement. The issue **Erich Hatala Matthes** raises in "Love in Spite Of," concerns loving someone whose practical identity includes properties that are bad for the beloved where the two requirements of love come into conflict. As Matthes goes on to explain, in addressing this tension, it is important to distinguish cases in which the bad property is a moral failing, such as having racist attitudes, and cases where the bad property is not a moral failing. In the first sort of case involving moral

failings with which the beloved identifies—as in Matthes' example of Racist Uncle who one loves in spite of his racist attitudes—one is morally justified in not fully satisfying the identity requirement. But in cases where the bad identity-forming feature is not morally suspect, in which, for instance, the harmful identity-forming feature is a disability, a different resolution between the two requirements on love is called for. Matthes here distinguishes between disabilities that involve social mediation (that is, ones that are due merely to social attitudes, such as body size) and disabilities that are "objectively" bad—bad independently of social attitudes. In cases involving a socially mediated disability possessed by the beloved, Matthes argues that there is no conflict between the identity and well-being requirements; in fact, loving someone partly because of such a disability can be a way of helping to undermine or at least mitigate the harm brought about by it. By contrast, in cases involving objectively bad disabilities (where there is also no antecedent moral reason to object to the identity-forming condition that is objectively bad for the person), Matthes argues that one ought to compromise the well-being requirement and embrace, as it were, the bad properties of the beloved. In such cases, to love the individual in spite of her disability would be deeply offensive to the beloved; the identity requirement in such cases trumps the well-being requirement. In explaining this verdict, Matthes distinguishes the property that is objectively bad for the beloved from the harm it causes. In loving someone partly in virtue of the property that is bad for her, one is not thereby loving the harm it causes, and so one remains in compliance with the well-being requirement.

In the final chapter, "Moral Reasoning," **Gilbert Harman** explores the complex nature of moral reasoning as an activity people engage in that can lead to a change in one's moral view. The particular model of such change that Harman embraces (whether construed as a normative model or as a descriptive model) is that of reflective equilibrium in which one attempts to find a balance between conservatism and coherence; between minimizing changes in one's view and reducing negative coherence on one hand, and enhancing positive coherence of one's view on the other. Harman proceeds to discuss various dimensions of such reasoned change in view. For instance, while in the theoretical realm, reflective equilibrium involves a (potential) reasoned change in one's beliefs that does not admit of wishful thinking or arbitrary choice, the same is not

true in reasoning about what to do. Moreover, as Harman understands reflective equilibrium in the practical realm, its inputs and outputs can be perceptions, feelings, and sensations, in addition to beliefs, desires, and intentions. A feeling of moral disgust at the prospect of engaging in some action can, for instance, be an input to one's reasoned change in view about whether to perform the action in question. Other dimensions of seeking reflective equilibrium include the extent to which reaching this state (or at least striving to do so) involves activity that is unconscious and thus implicit in one's reasoning. And, as Harman suggests, it might be that existing moral conventions in society are reached as a result of implicit social bargaining and adjustment—an instance of reasoning with others.

1

Taking Account of Character and Being an Accountable Person

STEPHEN DARWALL

I. RESPONSIBILITY, ATTRIBUTABILITY, ACCOUNTABILITY

Discussions of moral responsibility that follow in the wake of Strawson's "Freedom and Resentment" often note Gary Watson's distinction between "responsibility as attributability" and "responsibility as accountability" (Strawson 1968; Watson 1996).[1] For purposes of this essay, we can formulate this as the difference between attributing an action to a person by identifying elements of the person's character that gave rise to it and appraising them aretaically (as virtuous or vicious), on the one hand, and, on the other, holding the person accountable or answerable for the action, for example, by holding a reactive attitude like moral blame toward him.

Strawson was concerned with responsibility in the latter sense. A signature Strawsonian thesis is that there are distinctive states of mind—"reactive attitudes" such as resentment, guilt, indignation, and blame—through which we *hold* ourselves and others responsible in the sense of holding them answerable. Strawson called these "participant" attitudes because they are essentially relational or "inter-personal"; they presuppose "involvement or participation in a human relationship" (Strawson 1968: 79). Reactive attitudes are held from a perspective *within* relationship, and, at least implicitly, relate *to* their objects and make demands of them (Strawson 1968: 85).[2]

[1] I may be using this category of "responsibility as attributability" more broadly than Watson to refer to aretaic attribution more generally; his focus is more narrowly on what he calls "self-disclosure" views that relate actions to the agent's "practical identity."

[2] It will be important to keep in mind throughout that I am using "blame" to refer to the *attitude* of blame, whether or not it is expressed in *blaming*. Blame can exist as an attitude even if it is unexpressed: "I know she still blames me for it though she has never said anything to me

This relationship is often only implied and, in the case of "impersonal" reactive attitudes like "indignation" or moral blame, can be as thin as fellow member of the moral community. Even when we have no more particular relation to the object of our blame, Strawson holds, our blame implicitly relates *to* its object by implicitly making a demand of him or her (Strawson 1968: 87).[3]

Strawson and those who follow him hold that this "inter-personal" or *second-person standpoint*, as I call it, commits anyone occupying it to certain presuppositions regarding the powers and agency of those who are the objects of attitudes held from that point of view (Darwall 2006, 2013a, 2013b). Because "participant," second-personal attitudes have an element of implicit address, they are committed to certain presuppositions as what Gary Watson calls "constraints of moral address" (Watson 1987: 293–4). Intelligible address of any kind must assume that its object is capable of understanding and response; when we hold someone accountable we must presuppose that she has powers of moral agency that enable her to hold herself accountable as well.[4]

Participant attitudes contrast in this way with "objective," *third-personal* attitudes (Strawson 1968: 79). Objective attitudes do not carry the same presuppositions that are essential to attitudes from the second-person standpoint, even when they have the very same objects. Strawsonians typically hold, for example, that the attitude of blame presupposes that the object of blame was capable of knowing that what he did was wrong and of choosing not to do it for that reason, or for the reasons that made the action wrong. If these presuppositions are not met, then blame cannot be warranted, or perhaps even be intelligible.

But there clearly is no such problem with third-personal attitudes like being annoyed or being disgusted, even by some action someone performed. These are not "participant" or second-personal attitudes. If we come to believe that the object of our annoyance or disgust could not have known that her actions might appropriately give rise to these reactions, or that she could not modify her actions given this knowledge, this obviously has no tendency to show that her actions were not

about it." The Strawsonian thesis is that reactive attitudes, and not just their expressions to their objects, are implicitly "inter-personal" or second-personal.

[3] I am indebted to an anonymous referee for the Press for pressing me to clarify this.

[4] This is what I call "Pufendorf's Point" in Darwall 2006: 22–4.

genuinely annoying or disgusting; these attitudes might continue to be *fitting* responses to their objects (D'Arms and Jacobson 2000b). There might, of course, be reasons against being annoyed or disgusted by someone thus benighted or incapable. Perhaps responding in these ways is either unseemly or unfair. But these would not be reasons "of the right kind" to undermine either annoyance or disgust in its own terms (D'Arms and Jacobson 2000a; Rabinowicz and Ronnøw-Rasmussen 2004). The person's actions might remain just as annoying or disgusting, and so these responses might remain fitting in this sense.

This is not the case with blame and other reactive attitudes. Because they are implicitly addressed to their objects and make demands of them—for example, a demand to hold themselves accountable and take responsibility for what they have done—lacking the capacities to do this, *second-personal competence*, as I call it, tends to undermine blame in its own terms. It tends to show that the action was not really culpable. It either constitutes an excuse or, in extreme cases, may exclude the agent from the sphere of accountable moral agents who are even capable of being subject to or violating obligations and, therefore, of acting culpably.

Responsibility as attributability, by contrast, concerns itself with assessing action in relation to agents' *characters*, attempting to determine what motives or traits led to the action and how, therefore, the action bears on what Watson calls an "aretaic appraisal" of the agent, her virtues and vices. A paradigm example is Hume's treatment of "liberty and necessity" taken in conjunction with his virtue ethics.[5] "Actions are by their very nature temporary and perishing," Hume writes, "and where they proceed not from some cause in the characters and dispositions of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil" (Hume 2000: 2.3.2.6).

For Hume, moral judgments primarily concern motives (or motivated action) and character. Approbation and disapprobation, Hume's favored evaluative sentiments, always have some motive or trait of

[5] I mean to be offering Hume as a paradigm example of the broad category of aretaic attributability to which I referred above. Watson is mostly focused on self-disclosure or "deep self" views that require an element of autonomy and connection to the agent's "practical identity" that may not be present in many Humean cases.

character as their object in the first instance and only are transferred to acts by association.

> It is evident that, when we praise any actions, we regard only the motives that produced them, and consider the actions as signs or indications of certain principles in the mind and temper. The external performance has no merit. We must look within to find the moral quality. This we cannot do directly; and therefore fix our attention on actions, as on external signs. But these actions are still considered as signs; and the ultimate object of our praise and approbation is the motive that produced them. (Hume 2000: 3.2.1.2)

Praise or blame of actions, for Hume, therefore, is implicitly praise or blame for what motivated the actions, and it amounts to praise or blame of the agent only when those motives are appropriately characteristic of her, part of her character. Only then do they reflect on what Hume calls the agent's "merit," and render her "an object either of esteem and affection, or of hatred and contempt . . . impl[ying] either praise or blame" (Hume 1985: 173–4).

A major difference between responsibility as attributability and responsibility as accountability, then, is that the latter is primarily concerned with *actions* whereas the former is concerned with *character*, with characteristic motives. We generally hold agents accountable for what they do or do not do, and not, or not primarily anyway, for what they *are*. But praise and blame in the sense Hume has in mind is not primarily for what people do; it is for what they are as this is reflected in what they do.

My ultimate aim in this essay, however, will be to argue that although the primary focus case of accountability is intentional action, there are nonetheless ways in which we are accountable also for our character. Indeed, I shall argue, we are accountable for the trait of accountability itself—being disposed to hold ourselves accountable to one another by putting ourselves into second-personal relations of mutual answerability. I will proceed as follows. In section I, I explore the grounds of attributability and answerability's different foci—character and action, respectively—in fundamental differences in the attitudes they respectively involve. Section II concerns accountability's deep conceptual tie to the deontic (rather than the aretaic). I argue that the very concepts of moral obligation, right and wrong, cannot be understood independently of

accountability. Section III shows how attending to the difference between deontic and aretaic assessment can dispel the puzzle known as the "Knobe Effect," at least in the initial case in which Joshua Knobe discussed it: attributions of intentional action. Section IV then takes up the essay's positive argument and claim, namely, that despite accountability's primary focus on intentional *action*, we are nonetheless also accountable for aspects of our character. We are answerable for being accountable persons, for being disposed to place ourselves in second-personal relations of mutual answerability to others.

II. ATTRIBUTABILITY AND ACCOUNTABILITY: THIRD-PERSONAL AND SECOND-PERSONAL, RESPECTIVELY

There is a deeper difference that explains attributability and accountability's different foci on character and action, respectively. For Hume, praise and "blame" are essentially *aretaic*. They are appraisals of how good or bad a person is, where this assessment is made through sentiments of approbation and disapprobation, that is, *esteem* or disesteem, when we reflect on an agent's character or characteristic motivations from an observer's third-personal point of view. There is nothing essentially relational or second-personal, even implicitly, about them.[6] It follows that what Hume calls "blame" is not a Strawsonian reactive attitude. It is third-personal disesteem, as is shown both in Hume's claim that the distinction between "moral virtues" and "natural abilities" is only "verbal," and that being an "egregious blockhead" is a vice (see Darwall 2013c: 12–16).

Strawsonian accountability blame, by contrast, is second-personal and fundamentally *deontic*, as will become clearer presently. Reactive attitudes hold their objects to demands they presuppose are legitimate and bid, in second-personal relation, for their objects to acknowledge the legitimacy of the demands and the authority to be held accountable for acting in compliance with them (Strawson 1968; Darwall 2006, 2013a, 2013b).

[6] In, again, the logical or "grammatical" sense of not implicitly *addressing* their objects. Their objects may, of course, be people with whom we stand in relation. The point is that approbation and disapprobation do not implicit relate *to* their objects in the way that reactive attitudes do.

This means that fundamentally different attitudes are involved in ascribing responsibility as attributability and responsibility as accountability, respectively. We can put the difference this way. Attributions of virtue and vice concern how *estimable* someone is; they call on attitudes of esteem and disesteem. Assessing responsibility as attributability is thus assessing how an action should affect our esteem or disesteem of the agent. Blame as a reactive attitude, by contrast, is no form of disesteem. Whereas disesteem is third-personal, reactive attitudes like blame are second-personal "participant" attitudes through which we hold someone to a demand we take to be legitimate, bid for him to acknowledge the legitimacy of the demand, *take* responsibility, and hold himself to it.

Once we appreciate this fundamental difference, several observations follow. First, although the traditional contraries of praise and blame have a clear sense in responsibility as attributability, they do not in responsibility as accountability. Esteem and disesteem are contraries, and so are "praise" and "blame" as Hume uses these terms. However, accountability blame, understood as a second-personal holding-accountable *attitude*, has no true contrary. It might be thought that something like credit or merit is a likely candidate, but I would argue that the right way to think about blame as a holding-accountable attitude is not as according a kind of demerit to the person. That would place it too close to disesteem, which it clearly is not. And even if we distinguish between crediting and esteem as attitudes, maintaining that crediting (giving credit) is responsive to considerations of difficulty and effort in ways that esteem need not be, there is nothing essentially second-personal about the attitude (even implicitly) as a true contrary of blame would have to be.[7]

Another candidate for a contrary to blame might seem to be gratitude.[8] Unlike esteem and credit, gratitude is essentially second-personal, and Strawson explicitly categorizes it as a reactive attitude (Strawson 1968: 72). The problem is that gratitude is, like resentment, a *personal* reactive attitude, one that is felt from the perspective of a participant in the interactions to which it responds. Resentment is felt from the perspective of a victim of a wrongful injury, and gratitude is, at least

[7] I am indebted to Agnes Callard for discussion of this point. It is important to keep in mind that we are talking about blame as an attitude, rather than any act of blaming that might be taken to express the attitude.

[8] A number of people have made this suggestion to me when I have presented these ideas.

most typically, felt from the perspective of a beneficiary. Moral blame, on the other hand, is not held from such an interested position; it is held from a perspective of putative disinterest and is available to third parties. It is what Strawson calls an "impersonal" reactive attitude, but it is no less second-personal for that. Blame implicitly holds someone answerable from the perspective of a representative person or member of the moral community. So gratitude is not a true contrary to blame either.

I am not denying, of course, that a positive attitude like esteem can be expressed *to* someone, and that such an *expression* is second-personal, bids for uptake, and so on. Any such *communicative expression* is second-personal in this sense and carries the usual presuppositions about communicative address. We tend to reserve "praise" for such second-personal expressions of esteem.[9] So *praise* is second-personal in its nature. But it is not a positive analogue to the *attitude* of blame.[10]

Second, there may be reasons for being skeptical about judgments of overall estimability, and hence, for judgments of how a given action affects the agent's overall goodness or badness. Harman's and Doris' critiques, for example, put pressure on character attributions in general (Harman 1999; Doris 2005). And the idea that human agents can be arrayed on a continuum from good to evil may be problematic also for other reasons, as Peter Vranas has argued using psychological evidence suggesting that most people are capable of great good in some situations but also of great evil in others (Vranas 2005; see also Miller 2013). If profiles are sufficiently bivalent, overall character evaluations along a continuum may make little sense.

However, neither of these reasons for skepticism about the validity or helpfulness of character assessments and the degree to which specific actions reflect the agent's character has the same relevance to the

[9] One can hardly respond to the complaint that one never praised someone by saying that one did "in one's heart." There is no such thing as unexpressed praise. But blame can exist as an attitude even if it is never expressed. The Strawsonian point is that reactive *attitudes* are second-personal, not just that their expression to their objects is.

[10] "Humean" praise also takes motives and character as object and not just the intentional act, but not all forms of praise do. In particular, there can be forms of recognition that someone did her duty, maybe under difficult circumstances that bid for acknowledgment and uptake as in: "That was a 'stand up' thing to do." Perhaps we should not rule out the possibility of an attitude (solidarity?) that is itself second-personal that praise of this last sort might express. I am indebted to Ketan Ramakrishnan and Maggie O'Brien for discussion here. On solidarity, see Wiggins 2009.

question of whether an agent's action was culpable in the accountability sense. This latter question concerns whether the agent is fittingly held answerable for the action by herself and others through reactive attitudes like blame and guilt, and therefore whether she should take responsibility for having performed it. Whether this is so is simply a different question than how a given action reflects on someone's goodness as a person. One might be a complete skeptic about judgments of character while still taking accountability seriously.

Third, and especially importantly, clarity on this point is necessary to avoid forms of self-serving rationalization. It is possible, indeed depressingly common, for people to avoid taking responsibility for their culpable actions by saying to themselves that their record of moral action is sufficiently strong otherwise that the effect of the action in question on their overall goodness is negligible. (Like one bad grade on an otherwise unblemished transcript.) If the action is sufficiently "out of character" then it may not loom large in an overall aretaic assessment. But whether or not that is so is simply beside the point of whether the action in question was culpable (blameworthy in the accountability sense). For the agent to focus on the aretaic question may amount to evading rather than taking responsibility as accountability requires.

III. ACCOUNTABILITY AND THE DEONTIC

Having seen the difference between accountability and aretaic assessment, let us now take notice of accountability's essential connection to the deontic. In *The Second-Person Standpoint* and more recent work I have argued that accountability is conceptually linked to *deontic* (rather than to aretaic) concepts (Darwall 2006, 2013a, 2013b).

The deontic notions of duty, obligation, right, wrong, and moral permissibility can be defined in terms of one another. What it is morally obligatory or our moral duty is what it is impermissible and wrong not to do. And saying that an action is "right" can either mean that it is permissible ("all right"), or, when it is "the" right thing to do, that it is the only permissible action, therefore, morally obligatory and wrong not to do.

But what marks out moral deontic concepts in general? We invoke concepts like obligation and duty, right and wrong, when what we want to express is not just that there are reasons of a distinctively moral kind