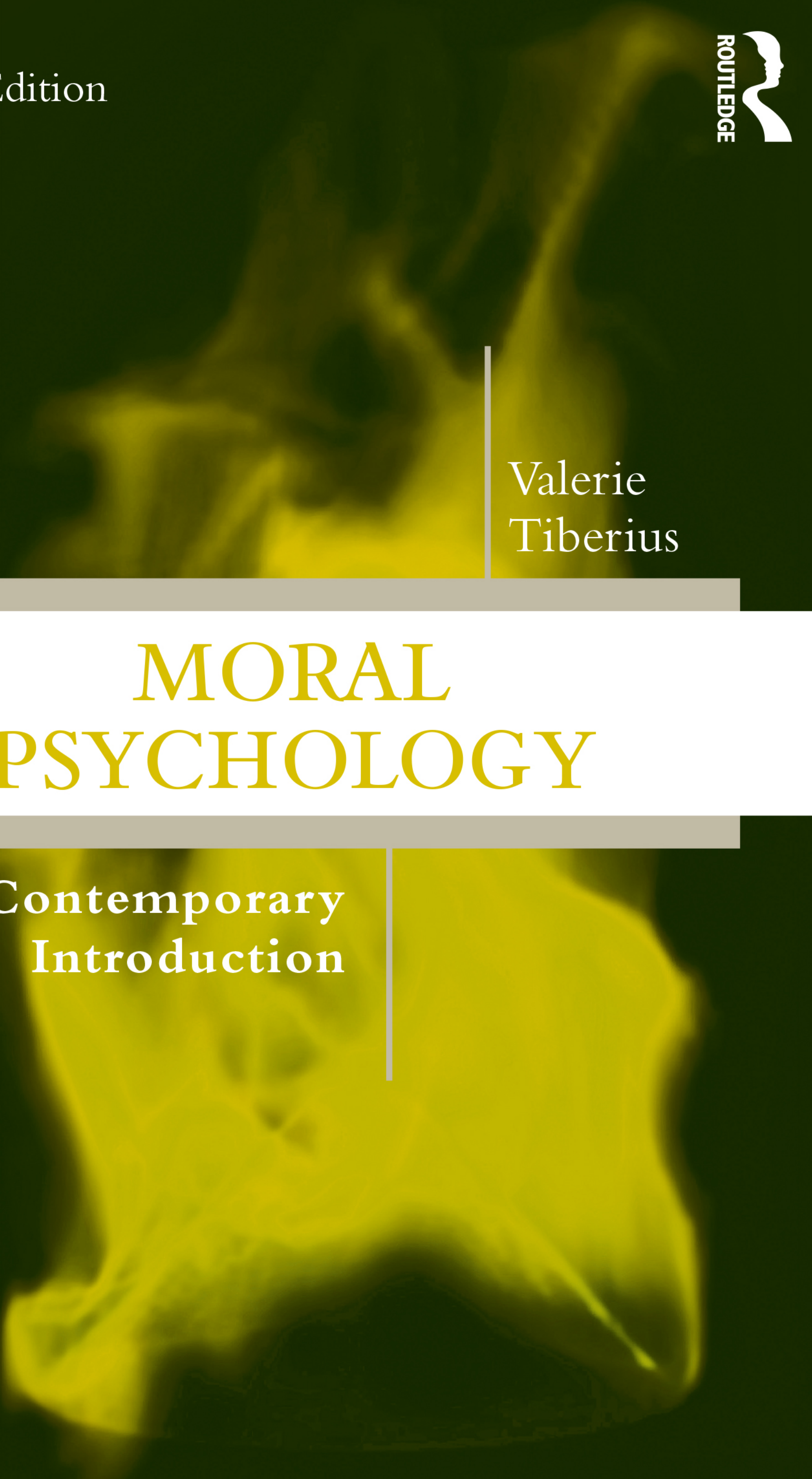
Second Edition



Valerie Tiberius

MORAL PSYCHOLOGY

A Contemporary Introduction

ROUTLEDGE CONTEMPORARY INTRODUCTIONS TO PHILOSOPHY

Praise for the First Edition

The best imaginable introduction to moral psychology. Unrivaled in clarity and scope, Tiberius offers a seamless and stimulating integration of cutting edge empirical research and foundational philosophical theories.

Jesse Prinz, *City University of New York Graduate Center*

Valerie Tiberius's book is an instant classic – the definitive introductory textbook on philosophical moral psychology. It is wide-ranging and authoritative, synthesizing an enormous amount of information, both scientific and philosophical. It's also a lot of fun. With her trademark wit and levelheadedness, Tiberius introduces readers to moral philosophy's most challenging problems and explains how recent scientific discoveries are casting these old problems in a new light. This book is beautifully written, filled with colorful examples that convey Tiberius' hard-won insights. Each of these chapters is fascinating in its own right, and together they tell a remarkable story that is just beginning. It's hard to imagine a better introduction to this exciting and rapidly advancing field.

Joshua Greene, *Harvard University*

This is the only text of which I'm aware – the very first one – on moral psychology. It addresses issues at the intersection of agency and normativity. It is a lively, well written,thoughtful book that is perfect for upper-division undergraduate courses. It is also a good book to bring graduate students and even scholars up to speed on the issues in this area. Highly recommended.

John Martin Fischer, *University of California, Riverside*

Several recent volumes offer collections of readings in the exciting interdisciplinary field of moral psychology, and lots more courses now cover the difficult questions arising at the intersection of philosophy (meta-ethics, moral theory, action theory) and psychology (and other mind sciences). But until now, there has been no text to introduce students to these questions or to guide them through these readings. Valerie Tiberius is an intelligent, clear, and engaging guide. In *Moral Psychology: A Contemporary Introduction*, she covers a wide range of topics, including moral motivation, virtue, well-being and responsibility, and she manages both to present competing views fairly and to take a stand on which views she favors. Students will learn a great deal from this book. I certainly did.

Eddy Nahmias, *Georgia State University*

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Moral Psychology

Released in 2014, this was the first philosophy textbook in moral psychology, introducing students to a range of philosophical topics and debates such as: what is moral motivation? Do reasons for action always depend on desires? Is emotion or reason at the heart of moral judgment? Under what conditions are people morally responsible? Are there self-interested reasons for people to be moral?

The Second Edition of *Moral Psychology: A Contemporary Introduction*, updates its responses to these questions, taking advantage of the explosion of recent research from philosophers and psychologists on these topics, and adding a chapter on the question of whether morality is innate or learned. As before, the book emphasizes the relationship between traditional and interdisciplinary approaches to moral psychology and aims to carefully explain how empirical research is (or is not) relevant to philosophical inquiry. The bulleted summaries, study questions, and lists for further readings at the end of each chapter have been updated.

Key Updates to the Second Edition:

- Includes a new opening section on human nature, borrowing material from elsewhere in the book
- Adds a new chapter on evolutionary and developmental arguments for the innateness of morality
- Expands coverage of the challenges to psychological research, including the replication crisis and the WEIRDness challenge
- Provides a new section on implicit bias and moral responsibility
- Offers enhanced clarity and accessibility throughout
- Includes up-to-date further reading sections and bibliography

Valerie Tiberius is Professor of Philosophy and the Paul W. Frenzel Chair in Liberal Arts at the University of Minnesota. Her previous books include – in addition to the First Edition of *Moral Psychology: A Contemporary Introduction* (Routledge, 2014) – *What Do You Want out of Life?: A Philosophical Guide to Figuring Out What Matters* (Princeton University Press, 2023), *Well-Being as Value Fulfilment: How We Can Help Each Other to Live Well* (Oxford University Press, 2018), and *The Reflective Life: Living Wisely with Our Limits* (Oxford University Press, 2008).

Routledge Contemporary Introductions to Philosophy
Series editor: Paul K. Moser, Loyola University of Chicago

This innovative, well-structured series is for students who have already done an introductory course in philosophy. Each book introduces a core general subject in contemporary philosophy and offers students an accessible but substantial transition from introductory to higher-level college work in that subject. The series is accessible to non-specialists and each book clearly motivates and expounds the problems and positions introduced. An orientating chapter briefly introduces its topic and reminds readers of any crucial material they need to have retained from a typical introductory course. Considerable attention is given to explaining the central philosophical problems of a subject and the main competing solutions and arguments for those solutions. The primary aim is to educate students in the main problems, positions, and arguments of contemporary philosophy rather than to convince students of a single position.

Recently published volumes:

Philosophy of Social Science
2nd Edition
Mark Risjord

Philosophy of Psychiatry
Sam Wilkinson

Philosophy of Emotion
Christine Tappolet

Ancient Philosophy
2nd Edition
Christopher Shields

Medieval Philosophy
Andrew W. Arlig

Moral Psychology
2nd Edition
Valerie Tiberius

For a full list of published Routledge Contemporary Introductions to Philosophy, please visit https://www.routledge.com/Routledge-Contemporary-Introductions-to-Philosophy/book-series/SE0111

Moral Psychology

A Contemporary Introduction

Second Edition

Valerie Tiberius

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Second edition published 2023
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge 2014

Library of Congress Cataloging-in-Publication Data
Names: Tiberius, Valerie, author.
Title: Moral psychology : a contemporary introduction / Valerie Tiberius.
Description: Second edition. | New York, NY : Routledge, 2023. |
Series: Routledge contemporary introductions to philosophy; vol 43 | Includes bibliographical references and index. |
Identifiers: LCCN 2022061211 | ISBN 9781032388465 (hardback) | ISBN 9781032388489 (paperback) | ISBN 9781003402787 (ebook)
Subjects: LCSH: Ethics--Psychological aspects.
Classification: LCC BJ45 .T553 2023 | DDC 170--dc23/eng/20230512
LC record available at https://lccn.loc.gov/2022061211

ISBN: 978-1-032-38848-9 (hbk)
ISBN: 978-1-032-38846-5 (pbk)
ISBN: 978-1-003-40278-7 (ebk)

DOI: 10.4324/9781003402787

Typeset in Times New Roman
by MPS Limited, Dehradun

Contents

Preface to the First Edition

When I was asked to write a textbook on moral psychology that included both traditional philosophical and new interdisciplinary approaches, I was excited, but also daunted. The field seems to me one of the most interesting and valuable areas of research in philosophy and the social sciences today, but it is also large and growing. No introduction could cover all the interesting work in one discipline, never mind more than one. Moreover, practitioners of philosophical and interdisciplinary moral psychology do not have the same conception of what the subject matter of moral psychology is, which makes it tricky to bring the two into conversation with each other. I think it is important, therefore, that the subtitle of the book is "A Contemporary Introduction." It is just that: an introduction, not the introduction. It is, furthermore, an opinionated introduction, like many of the other volumes in the Routledge Contemporary Introductions to Philosophy series. The way I have chosen to introduce the subject reflects my own interests and philosophical views. This would not, perhaps, be appropriate for a book that is a basic introduction, but I think it is the right approach for a book such as this one that is designed for advanced undergraduates, beginning graduate students and other academics with an interest in philosophy. A basic introduction that simply describes arguments without engaging in them would bore this intended audience.

One way in which the text reflects my own interests is that it includes a fair amount of meta-ethics. In part, this is because I think that psychological research has potentially important implications for meta-ethics. (I'm certainly not alone in this – some of the best known "new moral psychologists" work at the boundary between meta-ethics and empirical psychology.) I also wanted to write a textbook for an upper-level contemporary ethics course, where the philosophy instructor is interested in teaching moral psychology but for whom there doesn't yet exist an appropriate "moral psychology" course. I think this book is well suited for that purpose.

Another feature of the book is that is that it covers a wide range of topics and has therefore at times prioritized breadth over depth. In part, I made this choice because I want the book to be useful to people with a range of needs and interests. It seemed to me that since there is no other philosophy textbook on moral psychology at the moment, and since many philosophy departments do not offer a specialized moral psychology course, it makes sense to try to give an overview of the field that shows how it's connected to other topics in moral philosophy. I've also chosen to include some topics because of the way they catch people's attention (in the classroom and in the media), even where these topics might not be the ones analytic philosophers would deem most important. I think it's important to cover these topics and to encourage clear thinking about them, so that we don't get carried away by exaggerated pronouncements about what we know now about moral psychology. There is another motivation for breadth here, too, which is that when I learn about a new field I find the most difficult thing to do is to get "the big picture." To my mind, putting together a big picture is a useful task that can be performed by a textbook. My hope is that readers who want to delve into the details of a particular debate will find the big picture painted here to be good preparation for doing so. The lists of suggested readings at the ends of the chapters, and the cited works within the text, are a good place to start.

Though it is an opinionated introduction, I have tried to explain views with which I disagree carefully and charitably, and to consider how people with different views might take the debate forward. In my view, the virtue of charity in philosophical interpretation and argument is a crucial one that is under-rewarded in undergraduate philosophy (and elsewhere). The questions considered in this book are complex and difficult; it isn't going to be easy to answer them, and we are likely to do a better job if we are open to various perspectives and sources of information. I hope I have succeeded well enough in demonstrating open-mindedness and charitable interpretation that the textbook can serve as an example of these virtues, but, if not, then I hope students will be inspired by the magnitude of the issues to do better.

Many people helped me at various stages of writing this textbook. I would like to thank Michael Bishop, Fiery Cushman, Colin DeYoung, John Martin Fischer, Jesse Graham, Josh Greene, Claire Horisk, Eranda Jayawickreme, Victor Kumar, Joshua Knobe, Bob Kruger, Jimmy Lenman, Ron Mallon, Christian Miller, Eddy Nahmias, Shaun Nichols, Alexandra Plakias, Jesse Prinz, Connie Rosati, Adina Roskies, Chandra Sripada, Nina Strominger, Simine Vazire, Walter Sinnott-Armstrong, Stephen Stich, Karen Stohr, Liane Young, the Department of Philosophy at the University of Buffalo, the Moral Psychology Research Group and my Twin Cities Reading Group (Melanie Bowman, Jim Dawes, Michael Furstein, Daniel Groll, Stephanie Hake, Melissa Koenig, and Jason Marsh). I am very grateful to the National Endowment for the Humanities for a fellowship that supported me while I worked on this project. Thanks also to my editor, Andy Beck, in particular for having the idea that it was time for a textbook like this.

John Doris deserves special thanks. Had he not invited and encouraged me to join the Moral Psychology Research Group, I would not have been able to write this book and would not have met many of the inspiring researchers whose work is discussed here. I am especially grateful to Tim Schroeder for his painstaking and constructive comments on the first draft of the manuscript. Finally, as with all of my philosophical work, writing this book would not have been possible without the unflagging emotional and intellectual support of my husband, J. D. Walker.

Preface to the Second Edition

In the ten years that have passed since I began working on the first edition of this book, the field of moral psychology has exploded. As a quick illustration, Oxford published a handbook of moral psychology in 2010 that contains 13 chapters (Doris and the Moral Psychology Research Group). The new *Oxford Handbook of Moral Psychology* has 50 chapters and, according to the editors, it was difficult to limit it to that (Vargas and Doris 2022). If you can't cover everything in a thousand-plus-page multi-authored handbook, it certainly isn't possible to cover everything in a single volume like this one.

Instead of trying to do the impossible in the second edition of my own single-authored book, I have had to be selective. In the original version of this book, my goal was to draw connections between traditional philosophical questions and approaches and the newer empirically informed approaches to moral psychology. I still think this is a useful lens through which to view moral psychology. This may be most obviously so for philosophy instructors whose training was on the traditional side, or who would like to find a way to include empirically informed work into standard offerings of philosophy courses in ethics and metaethics. But I also think it's as true as it was ten years ago that this intersection of the traditional and the empirical is interesting for its own sake.

In revising the book, I had two main goals: to update the discussions of the various topics to reflect the current state of the field (at least, the part of the field this book stands in) and to make the book more accessible to undergraduate readers. Toward these aims, I have made minor edits in every chapter and I have updated references and suggested readings. I have added various new examples and a section on implicit bias and moral responsibility (in Chapter 10). In some places, I have changed terminology to match current usage. In Chapter 6, "Emotion and Moral Judgment," I made my taxonomy of emotion theories consistent with Christine Tappolet's new Routledge volume (2023) on the emotions. In Chapter 11, I replaced "hard determinism" with the more apt "hard incompatibilism" or "moral responsibility skepticism."

The second edition's structure and content are quite similar to the first edition's with a few exceptions. First, I have condensed what used to be the first two chapters into one introduction and I've eliminated entirely the discussion of evolutionary debunking arguments of moral realism. I thought the discussion veered too far into metaethics and went over like a lead balloon with students. The new introductory chapter has an improved discussion of philosophical methodology and a brief discussion of important criticisms of psychological research. It also addresses more directly a kind of naive moral subjectivism that I found to be a stumbling block for students while teaching this material.

Second, Part I of the book is now called "Human Nature: What Are We Like and What Does It Matter?" I have added a new chapter (Chapter 1) on the starting points of morality, which includes discussions of the evolution of morality and the development of morality in children. The chapter on the egoism–altruism debate is here (Chapter 3), with added discussions of Big God and principlism as alternatives to the altruism hypothesis. The chapter on the relationship between well-being and morality (formerly Chapter 10, now Chapter 4) closes Part I. To my mind, it made more sense to talk about the ways in which acting morally (or at least pro-socially) makes people happy immediately after discussing egoism and altruism. I think this better integrates the discussion of well-being into the narrative arc of the book, which might make it easier to work into a course plan. I can attest that students really enjoy discussing the topic of well-being and happiness!

Part II – "Moral Motivation and Moral Judgment" – contains the chapters on desires, reasons, emotions, moral judgment internalism and externalism, the debate about sentimentalism and rationalism, and virtue (previously Chapters 4, 5, 6 and 7). It now also includes the chapter on trolleys, "Brains, Biases, and Trolleys" (previously Chapter 11, now Chapter 8). As before, that chapter discusses questions about moral epistemology and the role of intuitions, but moving it to Part II allowed me to emphasize the important implications of "trolleyology" for questions about the nature of moral judgment.

Finally, when I wrote the first edition I was working out some thoughts about normativity and the relationship between is and ought, science and ethics. I am convinced that not all of these thoughts were helpful to readers, and so I have streamlined this theme. I removed the separate chapter on the is/ought gap and have confined myself to a few remarks about it in the conclusion of the book.

A Note about Pronouns

I stand proudly in the camp of people who favor a move to using “they” and “their” as gender neutral, singular pronouns. To me, this seems like the simplest and most elegant way to be more inclusive, given the options. So, where possible, I have moved to this usage in this book. Because of the kind of book it is, however, it hasn’t always been possible for me to do. Too often, I’m quoting other people and switching pronouns would be too confusing. And the fact that it is a second edition means that I have sometimes left things as they were. So, you’ll find the pronouns here a bit of a hodgepodge, which may not be a bad thing.

Acknowledgements

I continue to be grateful to all the people I thanked in the preface of the first edition of this book, but I now owe new debts of gratitude. Because the field has grown so much, it was even more important for me to rely on specialists with expertise on the topics that aren't in my own areas of research. I could not have written the new chapter on evolution and development without help from Victor Kumar, Pearl Han Li, Ben McMyler, and Josh Rottman, who read and commented on early drafts. I am grateful to John Doris, Jimmy Lenman, and J. D. Walker who also read chapters and provided invaluable feedback. My sincere thanks to the people who generously agreed to Zoom (or old-school telephone) interviews on their areas of expertise: in addition to Vic, Josh, and John, this includes Josh Greene, Ron Mallon, Eddy Nahmias, and Monique Wonderly. For indulging my pesky questions by email, I am grateful to Lara Aknin, Kiley Hamlin, Nathan Lackey, Adina Roskies, and Christine Tappolet (and again to Eddy, who deserves to be thanked twice). I also want to thank the many students in my "Philosophy of Psychology" class, who inspired me to improve the book, made helpful suggestions, and caught mistakes. Special thanks to the students who took the time to send written comments. And heartfelt thanks to the anonymous student who complemented my writing on the course website: "It is so nicely written and easy to follow, it felt like smooth butter on my eyes." That kept me going whenever I felt discouraged.

1 Introduction: What Is Moral Psychology?

Think about the last time you did a good thing. Maybe you helped a friend move, donated some money to a charitable organization, or took in a stray cat. What made you do it? Did you do it because you wanted to or because you thought you should? Are you just a good person? Did you think about a duty to help those in need? Were you thinking that you might want to ask your friend to help you move some day? Did the sad look on the cat's little face pull on your heart strings? And whatever the explanation, how did that happen? Are you a good person because of your upbringing? Or are your good impulses just an innate part of your human nature? Now think about the last time you did something bad. Perhaps you were in a hurry so you pretended not to see the cat or you broke your promise to help your friend move. Why did you do that? Are you just selfish? Were you overwhelmed by anger?

These are basic questions about moral psychology. They are questions about the psychological aspects of moral (or not so moral) actions. Questions about why we sometimes do the right thing quickly lead to other questions in moral psychology: is there a difference between doing something good and acting morally? Does it matter if we do something good but for the wrong reasons? Are only certain kinds of good deeds really praiseworthy? If so, which ones – actions done from duty, from virtue, or from sympathy? Are we really responsible for what we do? In the most general terms, moral psychology is the study of the psychological aspects of morality.

There are some ways of answering these questions that call on the expertise of scientists. If we want to know what was going on in your brain or your body when you saved the cat, we should ask a neuroscientist or a psychologist, not

DOI: 10.4324/9781003402787-1

a philosopher. But there are other ways of understanding these questions that explain philosophers' interest in them. Some of these questions involve concepts that philosophers study. For example, the question "Did you do it because you wanted to or because you think you should?" presupposes that *wanting* is different from *thinking you should*, and not all philosophers accept that this is true. And some of these questions are really not empirical questions at all. The question of whether only certain forms of moral motivation are good or praiseworthy is really a moral question, not an empirical one. Moreover, how the scientific questions are answered has important implications for what philosophers say about other topics in moral philosophy, and how scientists investigate these questions is very often influenced by their philosophical understanding of the phenomena. All this makes moral psychology profoundly interdisciplinary. Or at least it is today.

Moral psychology has changed dramatically in the last decade or so. Moral psychology in the 20th century focused on normative and conceptual questions and tended to dismiss the empirical questions. (It used to be common to hear philosophers say "oh, that's just an empirical question" in order to convey that it wasn't an appropriate topic for discussion.) It is fortunate that things have changed, because these different types of questions are so intertwined that it is very difficult to make progress on one set without making some assumptions about another. The way I understand *moral psychology* in this book does not exclude empirical questions and methods, and this reflects a growing consensus about how moral psychology ought to be done. Because I am primarily a moral philosopher (not a psychologist, cognitive scientist, or philosopher of science) this book is organized around questions at the intersection of moral psychology and philosophical ethics.[1] We'll be exploring how the more traditional questions of moral philosophy are related to questions that scientists are investigating, and how answering one can help answer another. Before we begin, it will be helpful to say more about the different questions and methods in moral philosophy and psychology.

Questions and Methods

In moral philosophy there are normative[2] questions, which are questions about what ought to be or what is good (such as the question of whether you only get any moral credit for what you do if you do it out of duty). There are conceptual or theoretical questions about what it makes the most sense to say about a given concept (such as the concept GOOD in the previous sentence.[3] And there are empirical questions about how to accurately describe the world that can be investigated by science (such as the question of what circumstances make people more likely to help strangers).[4] These questions are often all mixed up together. For example, consider this question: "Are people motivated to do what they morally ought to do?" To answer this question we need to know what it means to say that a person *ought* to do something (a conceptual or theoretical question). Once we know this, we also need to know something about what people ought to do (a normative question) in order to investigate what motivations people have to do it (an empirical question). From this simple example we can already see that empirically informed moral psychology and moral philosophy are profoundly intertwined.

Here's another example to illustrate the different kinds of questions we'll be talking about in this book. Consider the question "Should you be a vegan?" This is a normative question. It's a question about what you *ought* to do, not about what you are doing or what you're actually going to do. Personally, I think I probably should be a vegan, but I am not currently and probably will not actually become one (at least until the invention of delicious vegan cheese). Notice that it's also not a question about what people think. Most people do not think a vegan diet is morally required. If you want to know whether it is morally wrong to eat animal products, you are not asking whether most people *think* it's right; you are asking whether it *is* right. Normative questions are special, then, but to make progress answering them, notice that we'll have to answer many empirical questions, such as "Do cows, pigs, and chickens feel pain?", "Are animals caused pain by the way they are farmed for our use?", "Can humans survive without eating animal products?" We also need the answers to deep theoretical questions such as "Is pain the only thing that matters, morally speaking?" and "Does everyone's pain count equally?"

When I have asked my students about the difference between normative questions and empirical questions, one of the things they say is that empirical questions have factual answers that are not open to interpretation, while normative questions are subjective or relative and do not have definite answers. I don't think this is right. For one thing, many scientific findings are hotly debated and open to interpretation. For another, some ethical conclusions seem as factual as anything in science. I feel at least as confident about the normative claim that it's morally wrong to torture babies for fun as I do about the empirical claim that statins reduce the risk of heart disease. I think what people are reacting to when they think that normative ethical questions are "squishier" than scientific questions is that there doesn't seem to be a method for proving which answers to those ethical questions are correct in the way that there is in science. We need to investigate this alleged difference, but before we do there's one other kind of question to put on the table.

The assumptions people make about the difference between ethical claims and empirical claims raise questions about the *status* of the answers to our moral questions and of our moral theories themselves. Let's say someone tells you, "It's wrong to eat animals because animals are sentient beings." Is this a factual statement, like the statement "Animals feel pain"? Or is it an expression of an emotion, like "Boo! Don't eat animals!" Are moral statements such as "It's wrong to eat animals" the kinds of things that can be true or false? Are moral theories objective and universal? Such questions about meaning, truth conditions, and objectivity are *metaethical* questions. These metaethical questions are often thought to be conceptual. In the last century of analytic philosophy, conceptual analysis dominated the field and many analytic philosophers believed that philosophy is just the analysis of concepts. Typically, conceptual analysis proceeded by suggesting necessary and sufficient conditions for the application of a concept until a definition was reached that covered all the intuitive cases. Conceptual analysis has come under some fire recently. The analysis of concepts from the armchair (that is, without any empirical investigation) risks producing analyses that are idiosyncratic. Philosophers sitting in their offices might not use concepts (such as ought) in just the same way that everyone else does. If our goal is to characterize the concept as it is used by people in general, then the armchair method might not be a good one. Fortunately, the grip of the idea that "pure" conceptual analysis is all there is to philosophy has loosened recently. Now philosophers recognize that other methods and approaches are legitimate and can work together.

To answer these metaethical questions, then, philosophers use what we might call "theoretical analysis." We can think of this method as the method of figuring out what makes the most sense to say about some complex topic given all the relevant background information, including what we know from science, the purposes we have for developing a theory, and the constraints imposed by various theoretical virtues such as accuracy, consistency, scope, fruitfulness, and simplicity. Theoretical analysis may employ standard philosophical tools like conceptual analysis, counter-exampling, and thought experiments, but it also includes attention to broad theoretical goals and to what we know from science.

Notice that this kind of theoretical thinking is not only done by philosophers. Psychologists also construct theories, and the theoretical virtues I just mentioned are common to both fields.[5] Psychologists construct theories to explain their data, and the difference between a philosophical theory and a psychological theory probably has more to do with subject matter and type of data than with the nature of the theories. For some subjects, the two alleged types of theory may just be the same thing. One reason that psychologists need to employ theoretical analysis to refine the concepts they study is to ensure *construct validity*, which refers to the extent to which the way you have operationalized something so that you can measure it actually measures the very thing that you're interested in. For example, let's say you're a psychologist who wants to know whether rich people are happier than poor people. First, you need something you can measure. You devise a scale with some questions to ask people. To keep things simple, let's imagine that your scale just has one question: "How happy are you?" Then you get a random sample of people from the population, find out about their wealth, ask them your question, correlate the two variables – and voilà! Now you know! Or do you? The measure you have used has some serious construct validity problems. When we want to know how happy people are, do we really just want to know how they would answer this question? Probably not. We might want to know whether rich people are better off than poor people in some other way than just how they feel. (Do they get more of what they want? Do their lives have more rewarding or fulfilling experiences?) Or, even if we are just interested in how they feel, we might think that people's self-reports do not track how they actually feel very accurately. Before a psychologist does her research, she needs to ask what she is really interested in – how people say they feel, how people really feel, or something else altogether? In other words, she needs to define her concepts carefully before she moves on to empirical (scientific) investigation and this will involve her in theoretical analysis.

Of course, there are other methods we could use to inform our philosophical or our scientific research besides theoretical analysis. To determine how best to understand the target of investigation, we could employ the Dictionary Method. You want to know what happiness is? Merriam-Webster says it is "a pleasurable or satisfying experience." Done! But surely the controversy surrounding the nature of happiness that has persisted for thousands of years is not resolved so easily. Some people think it's mindful tranquility, some think it's the exercise of your human capacities, some think it's the ability to achieve your goals. Why should we listen to Merriam or Webster? To really probe the nature of happiness, we (philosophers *and* psychologists) need to engage in theoretical analysis. We need to consider why we want to know, what are the various options, and what are the pros and cons of the different theoretical options in terms of the virtues we're looking for in a good theory. Theoretical analysis isn't the only method for answering metaethical, conceptual, and other theoretical questions, but it is the best method.

Let's return to normative questions and the myth that there is no rational method for answering them. Against this myth, most moral philosophers would say that normative questions can be given better or worse answers by employing a specific kind of theoretical analysis. Moral philosophers interested in normative questions (typically) aim to develop theories that explain which actions are morally right and wrong, which states of affairs are good or bad, which traits of character are virtuous or vicious. In other words, they aim to develop theories that systematize and explain the moral considerations that guide us in life. They proceed by reflecting carefully on the implications of various possible principles or positions and refining their ideas until they arrive at a comprehensive and useful theory. Each moral theory has a different position on what kinds of considerations count as moral reasons and why. For example, questions about whether we have a moral reason not to eat animals or whether we have a moral reason always to tell the truth can be answered in a variety of ways. Utilitarians think that to answer such questions we should appeal to facts about pleasure and pain. Kantians think that we should appeal to considerations about rationality and respect. Virtue ethicists think we should appeal to the notion of human flourishing and the virtues that are necessary for it.

In normative ethics, there is a name for this special kind of theoretical analysis: reflective equilibrium.[6] This method is so central to moral philosophy that it's worth going into a little more detail about how it works. If you have taken a course in ethics, you probably recall that moral theories are often tested by how well they do at matching our intuitions. For example, most people have the intuition that it would be wrong for a doctor to kill a relatively healthy patient in order to harvest their organs to save the lives of four other people in the hospital. This example is used to criticize utilitarianism, because (so the argument goes) according to utilitarianism it would be morally right (in fact, required) to kill the one patient since more happiness is produced that way. Of course, there are many things the utilitarian can say in response to this example. The point here is just that moral intuitions are standardly taken to be relevant to evaluating moral theories. You might even notice this reliance on intuitions in your own decision making. Perhaps there has been a time when you have considered bending the truth a little bit in order to get ahead. You might have thought about whether an omission or an exaggeration of the truth is really a *lie* and compared it to cases where you have clear intuitions about the wrongness of lying.

Reflective equilibrium is a kind of theoretical analysis that emphasizes a particular kind of data, namely, our moral judgments or intuitions about specific cases, recognizable moral principles, and the theoretical virtue of coherence. At this point, it might occur to you to wonder why we should trust these moral judgments in the first place. If we shouldn't trust Merriam and Webster, why trust "us"? The short answer to this question is that we don't really have a choice. Our thoughts and feelings about what's morally right and wrong are the data that we start with, akin to our observations of the physical world that we rely on in science. But this comparison to observations in science immediately shows us that there must be more to this method than trusting our moral "observations". Sometimes our observations about the physical world mislead us. Living in the Midwest, it sure looks to me like the earth is flat, but I know this to be incorrect. If I don't currently have the moral intuition that it's wrong to eat eggs, couldn't I also be incorrect about that? To observe facts about microscopic particles, we need to have special equipment; none of our unassisted ordinary observations about bacteria or viruses are of much use. Couldn't things be similar in ethics? Might all of our moral judgments be incorrect because we lack a moral magnifying glass?

Reflective equilibrium can solve this problem in much the same way that science does: by taking account of more data from a wider range of sources and by allowing the best available theories to guide the inquiry. Astronomers do not just rely on how things look from the backyard. Virologists, thankfully, do not rely on ordinary intuitions about how viruses spread. Fortunately, moral thinkers also have more resources than our current sense of what's right and wrong. One thing we can do is to look at whether our intuitions are consistent across different kinds of cases. If you think it's wrong to eat bacon from a pig named "Babe," but not wrong to eat bacon from a nameless pig, something is probably wrong with your intuitions. If you think it's wrong to put chickens in cramped living conditions, but you have no moral concerns about eating factory farmed eggs, again, your intuitions can't all be correct. This sort of consistency reasoning allows us to clean up the data from our intuitions and is an important part of reflective equilibrium.

Some of the resources that improve reflective equilibrium come from moral psychology. For example, research on cognitive biases may inform the process of reflective equilibrium by revealing reasons to doubt some of our moral judgments about cases. We'll talk about the trolley problem as an example of this in detail in Chapter 8, but for a quick illustration of the point consider this famous example of framing bias: psychologists Amos Tversky and Daniel Kahneman presented people with scenarios in which a new disease is threatening 600 US citizens and various public health programs are proposed. In one case (problem 1), participants could choose between program A that will save 200 people or program B that has a one third probability of saving all 600, but a two thirds probability that no one will be saved. In the other case (problem 2), participants could choose between program C under which 400 people will die, or program D under which there is a one third probability that no one will die and a two thirds probability that 600 people will die. Most people choose program A in the first problem and most people choose program D in the second problem. This pattern of choices exhibits a bias. As Kahneman and Tversky explain,

> The preferences in problems 1 and 2 illustrate a common pattern: choices involving gains are often risk averse and choices involving losses are often risk taking. However, it is easy to see that the two problems are effectively identical. The only difference between them is that the outcomes are described in problem 1 by the number of lives saved and in problem 2 by the number of lives lost.
>
> (1981: 453)

We can imagine intuitions about choices among public health programs being used to make arguments in practical ethics: "It would be obviously morally wrong not to save 200 people for certain just on the chance that you could save more!" "Surely it would be unfair to let 400 people die for the sake of 200 others, when there's a chance we could save them all!" These two claims both sound plausible, but they pull in opposite directions. Evidence of biases shows us that we need to be careful about which intuitions we rely on. Understanding the psychological processes that underlie our intuitions may lead us to have reasons not to trust some of them, and acknowledging this will make reflective equilibrium better.

The discussion of methods in this section has focused on methods in moral philosophy. This is because people tend to be unsure of what philosophical methods are, whereas the basic scientific method is probably familiar to everyone: systematic observation, hypothesis formulation and testing, analysis and modification of hypotheses. The particular methods used by psychologists in moral psychology are too diverse for me to say anything very helpful about them in an overview, and we will see many of these methods in action throughout the book. It is worth noting at the outset, however, that psychological research has confronted a few significant challenges that may affect what kinds of philosophical implication this research has.

The first challenge, stemming from Joseph Henrich's 2010 paper "The weirdest people in the world?", is that the subjects of many studies in psychology are not representative of humanity in general. The subjects are, as Henrich explains it, WEIRD, that is, from Western, Educated, Industrialized, Rich and Democratic societies. Insofar as psychologists aim to draw conclusions about people in general – and insofar as philosophers look to psychological research to find solid claims about *human nature* – this could be a serious problem. Fortunately, researchers in moral psychology are wise to this problem now and tend to acknowledge it. Scientists and experimental philosophers are trying to diversify their samples, and people are better these days at reporting the demographics of their subjects. The best philosophers can do, it seems to me, is to be aware of this problem and on guard against making bad inferences from insufficient data.

The second challenge is the replication crisis, which threw a number of studies relevant to moral psychology under the proverbial bus. One of the core tenets of the scientific method is that other scientists ought to be able to repeat your experiment and confirm your results. In the last decade or so, psychologists have learned that some of their most famous and interesting experiments, when repeated, did not confirm the initial results. It's important to notice that (for the most part), those who have criticized studies for their lack of replicability and who have called for scientific reform are not accusing the scientists in the original studies of cheating or lying. There are a variety of reasons for the replication crisis that have to do with the norms in statistical analysis and publication practices; it's not a crisis caused by devious scientists with evil intent.

That said, it may still be a crisis, and if moral psychologists want to support their arguments and theories with reliable empirical work, they shouldn't ignore it. Here, again, I think the best that philosophers can do is to be cautious and aware. Edouard Machery and John Doris (2017) have a very helpful essay on best practices for philosophers interested in appealing to empirical research. One of their recommendations is to pay attention to widespread trends that have a body of research behind them, rather than picking on a small handful of studies. Do not cherry pick, in other words. We will talk more about the replication crisis in Chapter 9 where we discuss virtue and the situationist challenge.

I hope that the discussion of reflective equilibrium in this section has shown that the methods we use to answer normative questions are not so different from the methods we use to answer theoretical and scientific questions. In all cases, we look at the evidence we have (from our observations of the world or our intuitions about morality) and we aim to make sense of it in an accurate, consistent, and productive way. This isn't to say that there are no important differences between science and ethics; however, the difference is more subtle than the difference between "made-up crap" and "stone cold fact". There are better and worse things to say about morality, just as there are better and worse things to say about viruses, chickens, and human psychology.

How Are Psychology and Ethics Related?

Science and ethics have more in common than we may have thought, then, but this doesn't tell us how they are related to each other. Is empirical psychological research (our focus in this book) relevant to moral philosophy? This is itself a deep philosophical question: what is the relationship between the empirical facts, prescriptive ethical conclusions, and ethical theories?

There seem to be three possible answers to this question:

1. Empirical facts about our psychology have nothing to do with ethics.
2. Empirical facts about our psychology by themselves determine what is ethical.
3. Empirical facts about our psychology are relevant to ethics, but do not by themselves determine what is ethical.

We have already seen reasons to reject the first option. Research into cognitive biases is relevant to questions about how we ought to think about moral questions, and how we ought to revise our moral beliefs. Such research can help us construct better normative theories or come to better moral conclusions. We'll see many other examples throughout the book. For example, if virtues are difficult to cultivate without social support (something we learn from empirical investigation – see Chapter 9), then improving people's character demands a different strategy than relying on grit and willpower. Empirical research on how much rational control we have over how we behave seems to be relevant to normative questions about whether we ought to praise or blame people for what they do (see Chapters 10 and 11). For one more example, consider the widely held principle "ought implies can," which means that it can't be true that you *ought* to do something if you're completely unable to do it. If this is right, then our psychologies constrain what we ought to do. For instance, if "ought implies can" and if we are, in fact, only capable of acting for the sake of our own selfish interests, then it cannot be the case that we ought to act altruistically as some moral theories demand (more about this in Chapter 3).

What about the second answer? Do empirical facts dictate ethics? I belong to a long tradition in philosophy according to which you cannot derive an ought from an is.[7] To get from premises about the way the world is (including the facts about our own psychology) to conclusions about what ought to be the case, we're always going to need to introduce a normative or evaluative premise.[8] We could, for example, discover everything there is to discover about the psychology of our moral judgments about animal pain – what happens in the brain when we witness cruelty to animals; what sentiments, desires or beliefs are involved in making the judgment that people ought not to eat meat, and so on – and we would not have discovered whether it is actually wrong to eat animals. We could know all there is to know about the evolution and development of human morality – when it started, how it works, how universal it is – and we would not have enough information to draw conclusions about what is morally right and wrong. This is not an uncontroversial position, though it is a popular one among philosophers.

If the results of empirical science have something to do with ethics, but they do not by themselves determine what is ethical, then the third option must be the right one. Even if we can't derive an ought from an is, this does not mean that psychology is irrelevant to questions about what we have moral reason to do. The facts about the world – especially the facts about human psychology – shape how we ought to understand morality even if they do not directly yield conclusions about what's morally right and wrong. The chapters of this book elaborate and make a case for the third option.

What I've said is that there are three kinds of questions that are involved in moral philosophy: normative, theoretical (or conceptual), and empirical. These questions are often closely related in such a way that you must presuppose an answer to one in order to answer another, and you must answer more than one type in order to answer the big questions in ethics and metaethics. Of course, I've made things considerably more complicated than they were in the opening paragraph. We began by asking "Why do we act morally?" and "Why do we sometimes fail to act morally?" Let's return to one of these basic questions to see where we are. Why did you take in that stray cat? Notice that in taking this to be a question about moral action, we are assuming that taking in the stray cat was a morally good thing to do. This is a normative assumption that can be supported by a normative theory, which we would arrive at by applying the method of reflective equilibrium, which is itself informed by facts about psychology. Once that assumption is granted, we can propose some hypotheses for investigation about why you did it. Here are four:

- You wanted to.
- You felt sorry for the cat.
- You are a good person.
- You made the judgment that you have a duty to help suffering creatures when you can.

Thinking that these are four *competing* explanations assumes that these explanations are incompatible with each other. For example, it assumes that if you did it because you wanted to, then you did not do it because you felt sorry for the cat. It assumes that judgments about our moral duties are distinct from desires and feelings. It assumes that being a good person is different from wanting to help. The idea that these explanations are mutually incompatible depends on particular theoretical views about what desires, feelings, virtues, and judgments are. And defending those theoretical views requires both theoretical analysis and empirical research. Understanding the explanation of moral action, then, requires engaging with normative, conceptual, and empirical questions, and using all the methods at our disposal.

Structure and Aims of the Book

"What is moral action?," "Why do we act morally?," and "Why do we sometimes fail?" are deceptively simple questions. Indeed, one of the main aims of this book is to show you how complex these questions really are, and how much trying to answer them requires different methods of investigation and charitable interpretation of others' views. Philosophers have not always acknowledged that empirical research is relevant to their questions, and psychologists have not always acknowledged that normative and conceptual research is relevant to theirs. This has changed and the field of moral psychology has become much more interdisciplinary and collaborative, but there is still room for growth here. Another aim of the book, then, is to illustrate the potential benefits of acknowledging the mutual importance of the theoretical and empirical methods of inquiry. My take on these benefits, as I've said, occupies a Goldilocks position: empirical science is not irrelevant to ethics, but neither does it determine what is ethical. Rather, a middle position is the right one: The empirical study of moral psychology informs and constrains what we should say about ethics. Finally, the book aims to introduce key topics in moral psychology as practiced by philosophers for anyone who wants to understand what philosophers have said about these topics.

Part I of the book elaborates the Goldilocks position on various topics. We'll start, in Chapter 2, with a discussion of what we know from evolutionary science and developmental psychology about the raw material of our moral selves. Are we basically good? If so, in what way? We'll also discuss what it means that we are what we are. How does what we are like shape what we ought to do and how we should think about morality? We'll then turn to some very specific cases of psychological research into our moral "nature" and ask whether this research causes some problems for traditional views in ethics. Chapter 3 considers the challenge of psychological egoism. According to many traditional ideas about morality, motivation does not count as moral at all if it is self-interested. If all action is motivated purely by self-interest, as psychological egoism has it, then these traditional moral theories imply that moral motivation is psychologically impossible for us. There just wouldn't be any moral motivation according to Kant or Aristotle, for example, if psychological egoism were true. Chapter 4 asks whether the distinction between self-interest and morality is really as clear-cut as the debate about psychological egoism makes it seem. We'll consider the possibility that our nature might make it good for our individual well-being to be morally good.

Part II turns our attention to central debates about the role of desire, emotion and reason in moral judgment and moral motivation. Chapter 5 focuses on the thought that our desires are what explain why we sometimes act well and sometimes act badly. Here I will introduce the Humean Theory of Motivation (according to which desire is necessary for motivation and no belief can motivate us to do anything by itself) and the Humean Theory of Reasons (according to which having a desire to do something is a necessary condition for having a reason to do it). We will consider how well these theories are supported by an empirically informed moral psychology and what implications they have for what moral reasons we have.

Chapters 6 and 7 consider the possibility that moral judgments themselves are motivating, that is, that in judging something to be wrong we are *thereby* motivated to avoid doing it. This view takes two forms. In one, moral judgments are essentially expressions of emotions and, since emotions motivate us, making a moral judgment motivates us by itself. We consider this view, called sentimentalism, in Chapter 6 on emotion and moral judgment. The other form of the view that moral judgments are themselves motivating takes moral judgments to be rational judgments that motivate us to act morally insofar as we are rational beings. We consider this type of rationalism in Chapter 7. In the discussion of moral judgment in this chapter we will raise some problems for the idea that moral judgment is based entirely on sentiment, and we will consider what is at stake in the debate about whether moral judgments are based on sentiments or Reason.

Chapter 8 explores the idea that our moral judgments are the product of both emotional and rational processes. We'll consider one argument for thinking that some mental processes are more reliable than others in the moral domain. This argument draws on the trolley problem and research on what happens in the brain when people make moral judgments. The last chapter in Part II, Chapter 9, focuses on virtues: states of character that include emotional and rational dispositions that motivate us to act well. According to the virtue ethical tradition, moral motivation is motivation by virtuous character. We consider an argument from empirical psychological research that the kinds of virtues assumed by this tradition are very rare and likely not what motivates much of the behavior we ordinarily consider moral. As we'll see, the literature on virtues provides another excellent opportunity to look at the way in which empirical evidence is relevant to the assumptions made by moral philosophers.

Why people do what they do is relevant to another important set of questions in philosophical moral psychology. These are questions about praise, blame, and responsibility, the focus of Part III. In Chapter 10 we will consider what is distinctive about responsible agency: What is it about certain kinds of beings that makes it appropriate to hold them responsible for some of their actions? After a brief overview of the methodology that is used in debates about free will and responsibility, we consider two basic compatibilist positions. Compatibilism about moral responsibility is the view that people could be responsible even if determinism were true. In Chapter 11 we survey various arguments against compatibilism and we see how compatibilists have responded to these challenges. We also consider two different incompatibilist positions. This will lead us to investigate the methodology behind the free will debate in more detail and to ask what experimental philosophy has contributed to it. Finally, we will consider the claim made by some neuroscientists that investigating the brain proves there is no free will.

Summary

- Moral psychology is the study of the psychological aspects of morality. Two basic questions of moral psychology are, Why do we act morally? and Why do we sometimes fail? Answering these questions requires that we first figure out what counts as moral motivation for action. Another central question is, Under what conditions are we morally responsible for our actions?

- Moral philosophy in the broadest sense includes moral psychology, normative ethics and metaethics.
- There are three different types of questions in moral philosophy: normative, theoretical/conceptual and empirical/scientific. Philosophical and psychological research in moral psychology is relevant to all three kinds of questions.
- Theoretical analysis is particularly useful for answering normative and theoretical/conceptual questions. It is a method of careful reflection that aims to make sense of the target of investigation by attending to all the evidence, the relative merits of various theoretical proposals, and theoretical virtues such as consistency and fruitfulness.
- Reflective equilibrium is a specific kind of theoretical analysis used in the development of normative theories. Reflective equilibrium highlights the evidence of our moral judgments (or intuitions) about particular cases and the theoretical virtue of coherence.
- Normative, theoretical, and empirical questions are often related so that you can't answer one without assuming some answers to the others.
- There are three positions one could take on the question of how empirical science and ethics are related. This book rejects the two extreme answers and takes a middle position according to which the empirical facts are relevant to ethics, but do not by themselves determine what is moral.

Study Questions

1. As you begin reading this book, what questions about the psychological aspects of morality would you like to have answered? What "psychological aspects of morality" do you think are particularly important or interesting?
2. If you wanted to figure out whether a political candidate is a good person, how would you go about it? What kinds of questions would you need to ask? Would these questions be normative, conceptual or empirical – or some combination?
3. If you wanted to conduct a study of moral behavior with the ultimate goal of producing more of it in your community, how would you start?

Notes

[1] I use "moral philosophy" and "ethics" interchangeably throughout this book.

[2] In ethics, "normative" means having to do with what ought to be the case. It is opposed to "descriptive." It does not mean, as it does in other fields, "normal" or "typical." This terminology is so common in philosophy that I'm going to stick with it, even though it can be confusing to psychologists.

[3] It's a philosophical convention to use small caps when you want to talk about a concept, rather than the thing to which the concept refers. I'll follow that convention.

[4] For many philosophers, all questions are ultimately empirical in some sense. I will use the word *empirical* in a narrower sense that comports with the way philosophers often use it. *Empirical questions* are questions that can be answered by the methods of science.

[5] Indeed, this particular list of virtues comes from an online psychology textbook. https://nobaproject.com/modules/thinking-like-a-psychological-scientist

[6] Rawls (1951) is the classic defender of reflective equilibrium as a method for defending ethical theories. For more on this method, see Daniels (1979; 2008).

[7] The is/ought gap is most strongly associated with the Enlightenment philosopher David Hume (2000/1739: T3.1.1.27, p. 469).

[8] Note that one might have a metaethical theory according to which these normative premises themselves can be reduced to statements about matters of fact, but this would not make them any less evaluative and it would require sophisticated philosophical argument to justify the reduction. See Lutz and Lenman 2021.

Further Readings

Doris, John, Stephen Stich, Jonathan Phillips, & Lachlan Walmsley. Spring 2020. Moral psychology: Empirical approaches. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2020/entries/moral-psych-emp/

Nadelhoffer, T., E. Nahmias, & S. Nichols. 2010. *Moral Psychology: Historical and Contemporary Readings*. Wiley-Blackwell.

Sinnott-Armstrong, W. 2008–2014. *Moral Psychology*. Vols. 1–4. MIT Press.

Wallace, R. J. 2005. Moral psychology. In *The Oxford Handbook of Contemporary Philosophy*, F. Jackson & M. Smith (eds), pp. 86–114. Oxford University Press.

PART I
Human Nature: What Are We Like and What Does It Matter?

Philosophers have always thought that human nature and ethics are importantly related. They haven't agreed on the nature of that relationship, however. On one side, it's a beautiful marriage: to live ethically just is to live in accordance with your nature. On the other side, it's a bitter struggle: we must fight against our nature to live ethically. These very different positions share a common assumption, namely, that there is something that deserves to be called human nature – a way of being that we are all stuck with, no matter what. Further, they assume that what we're stuck with is pretty specific: it's either a good thing that gives us guidance for how to live, or it's a bad thing that tells us what to avoid.

As we'll see in the next chapter, this assumption is problematic. It's not clear that we are stuck with anything so specific. But we'll also see that we are probably not completely blank slates. There is a blueprint for human beings, even if that blueprint is vague and flexible. We'll explore what philosophers and scientists have said about our moral nature, the nature/nurture debate, and what it means for ethics.

One part of the blueprint that is hard to deny is that we are highly motivated by self-interest, which seems to pull us away from being morally good to others. In Chapter 3, we'll look at this specific debate about human nature – the egoism/altruism debate – and we'll consider its relevance to ethics. In Chapter 4, we'll explore the relationship between self-interest and morality more deeply, by investigating the nature of self-interest and individual well-being.

DOI: 10.4324/9781003402787-2

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2 The Starting Points of Morality

A pre-school teacher tired of listening to countless tales of woe from her 20 charges installed a "tattle-phone" in the classroom. The four-year-olds were eager to pick up the tattle phone and report on fellow students who weren't sharing, weren't helping, or who were generally being a bother. Fortunately, one of the fathers of these children was radio producer David Kestenbaum who replaced the fake phone with a phone that would record the children's voices. "Eli told a lie," said one child to the phone. "Seamus wasn't sharing with me, and I don't like it, and I'm very upset," said another. And – my personal favorite – "Nathan farted in my face, and I said, yuck, Nathan … And he didn't say excuse me."[1]

Young children have a sense of justice, it seems. They are aggrieved when they perceive others to do things that are bad: lying, hitting, stealing, and hoarding resources. Where does this come from? Do they learn it at day-care, or do they arrive at day-care already prepared to take offense when someone lies to them or hits them? This question about pre-schoolers scales up to the huge question – and the topic of this chapter – where does morality come from? Is it innate or learned, the product of nature or nurture, or a combination? This topic seems like a good place to start in a book about moral psychology. Whatever other topics we want to consider, it's good to know what the starting points are.

Before we dive in, it's worth pausing to ask why this debate might matter. There is a long tradition in philosophy (briefly mentioned in the introduction to this part of the book) according to which our nature *defines* what is ethical.

DOI: 10.4324/9781003402787-3

On this view, living in accordance with our nature is the ultimate ethical standard. If there is no human nature, or if human nature is bad, this tradition may be in some trouble. On the other side, entire ethical theories are built on the assumption that we are naturally selfish creatures who need to be restrained to act ethically.[2] On this view, our (selfish) nature impedes morality in ways that justify certain forms of enforcement of the moral rules. If there is no human nature, or if we are not naturally immoral, this tradition may need to be rethought. On both sides, the assumption is that there is some essential human nature that is true for all human beings, no matter what culture they inhabit. This universal, essential nature informs moral theories that are meant to be true for everyone for all time. From the point of view of many ethical theories, then, it matters whether there is such a thing as human nature and, if there is, whether it is good or bad. We'll discuss why this debate matters in more detail at the end of the chapter; for now let's return to the question of whether morality is innate.

To even begin to answer this question, the first thing we'll need to do is define our terms. Let's start with the term "morality." When I just now googled "morality," I got this: "principles concerning the distinction between right and wrong or good and bad behavior." That's not bad – certainly morality concerns what is right or wrong and good or bad. But is morality only about principles? What about moral emotions and intuitions, moral virtues and vices? Should our definition be more capacious, perhaps? Are the distinctions between right and wrong or good and bad always *moral*? What about things that are aesthetically good or bad? Should our definition be more specific? And, if we can't just rely on the dictionary, how do we even try to define morality in the first place?

We can distinguish two methods for defining morality. First, we can make theoretical or conceptual arguments, or, second, we can look and see how morality is practiced by human beings. On the conceptual side, Immanuel Kant argued that it is part of the very idea of moral duties that they are categorical, that is, that they apply to us independently of how we feel or what we want. If it is morally wrong to lie, he thought, then it's wrong to lie no matter whether you would benefit from lying, or whether the pros outweigh the cons. Many philosophers have agreed with Kant's approach, even if they haven't agreed with all his conclusions. Psychologists like Lawrence Kohlberg and Judith Smetana have followed in Kant's footsteps. For example, Smetana and colleagues (2013), citing philosophers, say that

> moral concepts pertain to forms of social interaction that are universally applied (that is, to everyone) and obligatory, impersonal (in that they do not depend on personal preferences), and based on their intrinsic features, such as their consequences for others' rights and welfare.
>
> (p. 24)

On this view, held by some philosophers and psychologists, moral judgments are, by definition, universal and categorical.

It shouldn't be surprising that experts who hold this view about morality have not been the ones talking about the innateness of morality. For them, moral judgments are rather sophisticated: understanding universally applicable rules and norms that apply independently of inclination is an intellectual achievement that has to be learned. Indeed, Kohlberg's (1984) theory of moral development, which was highly influential in the middle of the last century, holds that there are moral stages that children proceed through as they develop cognitively. These stages culminate in a kind of moral thinking that would have made Kant very happy because of its emphasis on universal rules.

We will return in later chapters to Kant and to this theoretical approach to the question about the nature of morality. For now, we are going to focus on the other approach, which is to look at human morality and see what it's like. There's a certain imprecision in this approach: how do we know we're observing *morality* if we don't have a definition of morality? I'm going to suggest that we try not to worry about this yet. It will be sufficient right now to operate with a rough intuitive sense of morality – a "you know it when you see it" grasp of the concept – in order to pick out *moral* emotions, *moral* norms, and *moral* capacities. But before we get there, there is another term in our question that needs to be defined: "innate."

The term "innate" is problematic in various ways and we may want to avoid it, or at least be very careful about what we mean by it. The problem with the term "innate" is that many people assume it means "hard-wired" or "unchangeable" so that what it means to say that a person is innately kind, for example, is that they will grow up to be kind no matter what kinds of experiences they have. No expert thinks that morality is innate in this way. Instead, experts who say we have "innate" traits think that we have certain tendencies or dispositions prior to experience that are profoundly shaped by our environments. They think of "innate" moral capacities as a "rough draft" that is edited by culture and upbringing (Graham et al. 2013), or as the "starting points" that shape (but do not determine) development (Rottman & Young 2015).

Once we understand what we're asking when we ask whether morality is innate, three other big questions arise. First, how do we answer this question? What kind of evidence do we need to show that something is a starting point? Second, what specific capacities or dispositions do we have prior to experience? What's in the rough draft? Is the four-year-old's sense of fairness in there? How detailed is the draft? And third, what does any of this matter for moral philosophy?

Drafts, Starting Points, and Taste Buds

It's not unusual to hear people from all walks of life engaging in the nature versus nurture debate. Topics that have to do with gender often provoke this debate. Perhaps, like me, you have heard parents of young children talk about the preferences of their toddlers. "We treat our son and daughter exactly the same, but he likes trucks and she likes dolls! – it must be genetic." You have probably heard the odious claim that "boys will be boys," which seems intended to excuse aggressive (sometimes violent) male behavior by pointing out that the aggression is part of boys' nature. I've even heard philosophers say that the reason there aren't more women in philosophy is that women are not by nature argumentative (apparently ignoring the number of women in law). The non-scientific conception of nature versus nurture seems to assume that whatever is natural is fixed, because it is "hard-wired" into our psychologies before birth.

This folk debate has two poles: on one side (the side just mentioned) we have traits that are natural, innate, hard-wired, and fixed. On the other side, the human mind is a blank slate and all the traits we end up with are learned from experience. Scientists and philosophers tend not to take these extreme positions. Instead, experts recognize the importance of interaction between nature and nurture. Almost everyone these days thinks that *something* is given by nature, but that what's given is not a precise script that will be followed no matter what happens. The real debates here are about exactly what is given by nature and how that genetic endowment interacts with environmental factors.

We might use the word "innate" to describe this natural endowment, but the problem with that word, as I mentioned briefly above, is that it is very strongly associated with a *lack* of environmental influence. In research on the folk concept of innateness, Paul Griffiths, Edouard Machery, and Stefan Linquist (2009) looked at how ordinary people (what they call the "biologically naïve") use the word "innate" and one of their findings was that people think of innate traits as fixed, by which they mean that "the trait is hard to change; its development is insensitive to environmental inputs in development" (2009: 609). If the truth is that our psychological traits are the product of genes and environment working together, this assumption of fixity could really lead us astray. This is true in the moral context. For example, if you accept that "ought implies can" (the very popular idea that you can't be obligated to do something it is impossible for you to do) then what we are actually capable of constrains what we morally ought to do and our views about what is unchangeably fixed into our psyches will limit what we think is morally possible. This could have very bad effects – consider, for instance, the consequences of thinking that boys are unalterably incapable of controlling their aggressive impulses. Because the ordinary term "innate" has such strong assumptions built into it, Griffiths and his colleagues recommend abandoning it and using a more precise term when you want to talk about the nature side of the debate.[3]

What is the best more precise term for our purposes? If we look at how philosophers and psychologists talk about "human nature" or "innateness" when they are asking about our moral capacities, the key concept seems to be that "innate" traits are *unlearned*. We do not gain these traits – such as sympathy, or a sense of fairness, according to some – by learning from experience. This fits well with the way that the psychologists Jesse Graham and Jonathan Haidt (whose research we will come to shortly) define the word "innate." They specify that what they mean is "organized in advance of experience" and they go on to spend considerable time talking about the importance of interaction with the environment. For Graham and Haidt, what is "organized in advance of experience" is like a rough draft of our moral psychology, which gets edited as we learn.

Graham and Haidt sometimes express their idea in terms of our "moral taste buds." These are the capacities that enable and shape our moral experience, in the way that our taste buds shape our experience of food. We do not have to learn to use our taste buds, but how they develop and shape our food preferences is profoundly influenced by (among other things) what foods we experience in our families and cultures. Similarly for morality, the idea is that we are born with certain moral capacities that are then shaped by our cultures, families, and experiences. Developmental psychologists Joshua Rottman and Liane Young talk about moral "starting points," which are where we start *before* we have had a chance to learn anything. Taste buds, starting points, and rough drafts are all getting at the same idea: there is something we have that is unlearned.

"Unlearned" is not a very familiar word, so it will be easier for us to remember that it does not mean hard-wired or impervious to environmental influences. That's an advantage. Does it also help us draw a connection to the traditional concerns of moral philosophy? I think so. Unlearned traits are likely to be very widely distributed across the human population, because they are what we have *before* our particular culture exerts its influence. They also make good candidates for inputs to moral theory, because – even if they are quite flexible – they do give us some raw material to work with. Defining "innate" as "unlearned," then, is a good way to go because it is both compatible with how scientists talk about our moral natures and of potential relevance to moral philosophers. In the rest of this chapter, we're going to think about what moral capacities we human beings have in advance of experience. Are there moral taste buds or starting points? And if there are, what are they and what bearing do they have on ethics? Because of the misleading connotations of the word "innate," I'm going to follow the suggestion to avoid the word when possible (sometimes I have to use it, because other people do!). My experience in teaching leads me to think it's quite difficult to get people to upgrade their ordinary concepts to something more precise. So, instead, I'll use the word "unlearned" and I'll talk about rough drafts, starting points, and taste buds. The view that there are moral dispositions prior to learning is often called "moral nativism" and I'll use that label here, too.

What's the Evidence?

One major source of evidence for moral nativism brings us back to the tattle phone: early emergence. The earlier in childhood that a trait (an emotional, cognitive, or behavioral disposition), can be detected, the more likely it is that this trait is not learned. This makes sense because newborn babies have had almost no experience of the world, no time to learn; pretty much everything they do when they first enter the world is something organized prior to experience. A brand new baby will grasp your finger if you put it in their hand and this grasping behavior is a reflex that happens without your having to teach them. Grasping is native to babies. Is morality?

Research confirms that pre-school kids (3–5-year olds) have some morally significant traits. They have a sense of fairness, as illustrated by the tattle phone. They also have sympathy for the suffering of others and a sense that it is wrong to harm other creatures. They even prefer to play with helpful people and think that people who harm or hinder others ought to be punished (Van de Vondervoort & Hamlin 2017).

The problem with 3–5-year olds, however, is that they have had some time to learn about fairness at home or from prior experiences at school. For this reason, developmental psychologists interested in exploring moral nativism have had to look at much younger children who have had even less experience with the world from which to learn about morality. Research on very young children is, as you can imagine, tricky. You can't *ask* a 6-month-old child whether she thinks someone is good or bad or which person she thinks should be punished. You also have to worry about "fussing out" – a very cute term for annoying behavior that I learned from developmental psychologists. Babies fuss out – become too disruptive or distracted to continue participating in the study – at a much higher rate than older kids or adults. Psychologists who study 6–10-month-old children must have an impressive amount of patience and perseverance.

What has this research on babies found? In a landmark study that has been highly influential, Kiley Hamlin, Karen Wynn, and Paul Bloom (2007) found that 6- and 10-month-olds prefer helpers to neutral agents and to hinderers, and prefer neutral agents to hinderers. You might be wondering how they assessed the preferences of 6-month-olds who cannot express their preferences in words. Even though babies don't have command of the English language (or any other language for that matter), they do have ways of expressing themselves. Reaching is particularly important in this research – babies reach for things they want and what they reach for can be observed and recorded.[4]

To test babies' preferences, Hamlin and her colleagues first showed them some videos of adorable shapes with large eyes – a square, a circle, and a triangle in primary colors – climbing a hill. One of the shapes – let's say a red circle, though they varied the shapes and colors to make sure they weren't just picking up on the babies' preferences for those things – tries unsuccessfully to get up the hill. Then a new shape enters the scene! Blue Square pushes Red farther down the hill, hindering its pursuit. In the next scenario, Yellow Triangle pushes Red up the hill, helping it to fulfill its goal. After the babies have watched this until they are bored, the experimenters present them with a choice between Blue Square (the hinderer) and Yellow Triangle (the helper) and encourage them to reach for one. In the earliest experiment, 14 out of 16 10-month-olds and all 12 of the 6-month-olds reached for the helper. Pretty clearly not just the result of chance. Hamlin and her colleagues conclude from this study and many others like it that the infants prefer the helper.

There are reasonable criticisms of this research. You might worry that there are other explanations of the infants' behavior. For example, perhaps the babies just have a preference for things moving up – a kind of visual preference that has nothing to do with their perceptions of the agency of the shapes. This is a possibility that the researchers who did these studies thought of at the time. To rule it out, they tried the same experiment with cute little Red replaced by an inanimate red circle with no eyes. This red circle was just presented as a block with no personality and no goals. When babies watched the same scenes with this version of the red circle, they chose Blue Square or Yellow Triangle at chance, confirming the idea that the babies are responding to *helping* and *hindering* Red. When Red wasn't the kind of thing that could be helped or hindered, they responded quite differently. Of course, this is just one alternative explanation of the results Hamlin found. Various alternative explanations have been considered and explored, and the results of these experiments have been confirmed and extended many times by other researchers.[5]

Hamlin and her colleagues interpret their findings as evidence

> that preverbal infants assess individuals on the basis of their behaviour towards others. This capacity may serve as the foundation for moral thought and action, and its early developmental emergence supports the view that social evaluation is a biological adaptation.
>
> (p. 557)

In other words, they take the evidence from research with infants to point to the existence of moral starting points, pathways along which we will tend to develop that are organized prior to experience. These moral nativists do not believe that the starting points determine what develops regardless of experience. Most moral nativists these days agree that cultural learning profoundly shapes how the moral starting points end up.[6] Science is always a work in progress, but my own assessment of the evidence is that the burden of proof has shifted onto those who deny that there is something about morality that is prior to learning. Exactly what that is will be the question for the next section. For now, let's turn to the other main sources of evidence.

In addition to developmental evidence, defenders of nativism look to evidence from evolution. To explain why a moral trait would have emerged prior to experience, nativists cite the fact that the trait conferred an advantage in human evolution. The idea here is that evolution wrote the rough draft, so we should be able to see why it would have written it the way it did (don't let my metaphorical way of talking allow you to assume that evolution has its own intentions and goals: it doesn't). Now, evolutionary explanations are not sufficient evidence for nativism, but they are an important piece of the puzzle.

The basic consensus among people who think about morality as an evolved human practice is that groups of humans who were able to cooperate with each other, did better than humans who didn't play well with others. Morality – or at least some basic moral capacities like sympathy for the suffering of others and a sense of fairness – helped us get along. Morality inclined us to take care of each other and to cooperate to procure food and secure our safety. Eventually, morality allowed us to live (relatively) peaceably in large groups that brought even greater evolutionary advantages.

This idea that morality solves a social problem is actually a very old philosophical idea that predates the understanding of evolution. Thomas Hobbes (1588–1679), for example, argued that moral rules, dictated by an all-powerful sovereign, are what save us from the short and nasty lives we would have in the state of nature. The state of nature, for Hobbes, is our natural state – the way we would be without a ruler to enforce the rules. Because we are (he thought), by nature, selfish, violent, and short-sighted, our lives would be much worse in the state of nature than they could be if we all agreed to consent to a ruler who would keep us all in line. Hobbes' view about human nature seems pretty pessimistic, in light of what we know about children. It also just seems incorrect in light of what we know about evolution. Nevertheless, the idea that the point of morality is to help us live better together is not an unfamiliar novelty.

Research on evolution changes Hobbes' simple picture. First, there was no specific point at which we moved from pre-moral to post-moral. (Nor is it certain that Hobbes thought his story was historically accurate – there are different interpretations of social contract theories on this point.) Evolution is a process. Second, evolution changed our nature. The way Hobbes sees it, once we are in civil society, we are the same selfish creatures, but because of the threat of punishment it pays to follow the rules. The way moral nativists see it, evolution changed us from creatures with limited moral emotions and capacities, to creatures with much more sophisticated moral emotions and capacities. In particular, through the process of evolution, the basic sympathy and loyalty we share with apes expanded to a much broader sympathy that extends even to those who are not our kin. We also developed a sense of fairness, the capacity to articulate and follow rules, and the capacity to reason together about what rules we ought to have.

Let's look at a few of these evolved moral capacities more closely. It's not possible to tell the whole story of the evolution of morality here (and I highly recommend the readings on this topic at the end of the chapter). For our purposes, it will suffice to consider some examples of the kinds of arguments nativists make. First, consider altruism and kin selection. You might think that altruistic desires (ultimate or non-instrumental desires for the good of another) make no sense from the standpoint of evolution, because my desire for another person's well-being is not going to make it more likely for me to survive and, indeed, might make my survival less likely if it inclines me to sacrifice my interests for the other person before I get a chance to reproduce. But, as Eliot Sober and David Sloan Wilson (1998) have argued, altruistic desires for our children's well-being (and the well-being of other members of the clan upon which we depend) do make a good deal of sense from the standpoint of evolution. Sober and Wilson argue that altruistic ultimate desires are a more reliable way to ensure that human beings would take care of their children than counting on ultimate desires for the parent's own pleasure that then give rise to instrumental desires for children's well-being. Altruistic desires for our children's well-being would provide direct motivation for us to care for them that would not be undermined by our finding easier ways of getting pleasure. If altruism is a better, more reliable way to get us to care for our children than selfish motives, then altruism would have been adaptive and could have been selected for evolution. (Note that this same form of argument could be used to argue that some non-human animals have altruistic desires, sympathy and/or compassion; see Andrews & Gruen 2014.)

The second example of the evolution of morality is cultural group selection for the capacities to understand, internalize, and follow rules or norms. The idea of cultural group selection is that a culture can be selected for in evolution if it outperforms a different culture. It might win the competition by being better in battle, which it might do if it has greater trust among members and enough organization to facilitate coordinated fighting. Or, it might win the competition by having more attractive norms and values that incline other groups to copy it, which it might do if it highlights values that make life easier, like caring and sharing. In their book *A Better Ape*, Victor Kumar and Richard Campbell argue that

> Human groups with norms were more successful in intergroup competition than human groups without them. Norms endow a group with greater cooperative ability, especially in response to new ecological challenges. For example, more cooperative groups could reliably find food – and avoid becoming food. They also gained an edge in cultural evolution by pooling their information resources. Humans that are generous with new ideas will tend to form groups that are more successful than humans who are stingy with new ideas.
>
> (Kumar & Campbell 2022: 78)

Economists Samuel Bowles and Herbert Gintis argue that the capacity to *internalize* the rules was also selected for in cultural evolution:

> Groups that created institutions to protect the civic-minded from exploitation by the selfish flourished and prevailed in conflicts with less cooperative groups. Key to this process was the evolution of social emotions such as shame and guilt, and our capacity to internalize social norms so that acting ethically became a personal goal rather than simply a prudent way to avoid punishment.
>
> (Bowles & Gintis 2011; see also Richerson & Boyd 2008)

These ideas are reflected in the ordinary experience of working in a group or a team. Teams work better together when there are some agreed upon rules for how to proceed, to which everyone is committed: who will do what and by what deadline, how will information be communicated and on what schedule. I am not suggesting that successful teams these days will produce more children – the stakes have changed since we were hunters and gatherers. My point is just that we can glimpse from our own experience the dynamics that may have created an evolutionary advantage in our history.

We have now moved a fair distance away from the basic unlearned moral capacities. The research here does not show that specific moral rules, reasoning practices, or enforcement mechanisms are the product of evolution. Rather, the claim is that there are certain basic emotional, motivational, and cognitive capacities that evolved in human beings because these capacities gave us or our culture an advantage. The complex and culturally specific moral behaviors and practices we have today are learned, but they have their roots in evolved capacities. Of course, the evolutionary explanation of the existence of these capacities isn't proof that these capacities are unlearned. As psychologist Audun Dahl and colleagues observe, "Of course, evolutionary benefit could at most be a necessary criterion for nativism, and never a sufficient one. Even characteristics that are highly valuable for survival, such as wariness of heights, may depend on specific experiences, such as self-produced locomotion" (2021: 3). Nevertheless, if we do have unlearned moral tendencies, it would be surprising if there were no evolutionary explanation for them. It would start to look more plausible that these tendencies are learned, albeit very early in development – and that would be bad news for nativism. So the evolutionary evidence is an important piece of the puzzle.

In this section we have talked about evidence for nativism from developmental psychology and the study of the evolution of morality in humans. Some researchers also look to evidence from research in primatology. If our moral capacities evolved to be present in human beings prior to experience and learning, we should expect to see similar traits in our closest living relatives, non-human primates. The thought here is that up until 4–6 million years ago, our evolutionary history was shared with chimpanzees (and shared with gorillas until 7–9 million years ago). Even though this seems like a long period of time, in the grand scheme of things it is not very long at all, which means that we should expect to have a lot in common with the other great apes. Finding moral capacities that are similar to ours in these creatures would support the claim that these capacities are evolved starting points. Though I won't discuss the details here, it turns out that primates (and indeed some other mammals) do have capacities that are analogous to some of our moral capacities, such as loyalty and sympathy with respect to their kin group (De Waal 1997; Hrdy 2009).

What's in the Draft?

Inspired by thoughts about *why* morality evolved, moral psychologists have been investigating what *sort* of morality evolved. One of the most influential research programs on this topic in moral psychology started with Jonathan Haidt's Moral Foundations Theory (MFT), so we'll begin there too.

Together with Jesse Graham and other colleagues, Haidt proposes that there are at least five moral domains or taste buds. Each taste bud, they argue, is a response to a different adaptive challenge that our human ancestors faced. Each taste bud engages different moral emotions and tends to be associated with different moral virtues and vices and different moral norms or rules. Here are the five contenders from MFT:

1. Care/harm: We feel compassion for victims of suffering and anger at those who inflict suffering. It was adaptive for us to have these emotional tendencies because we needed to care for our vulnerable children. Now, our compassion is engaged by a wide range of triggers, including cute baby animals. The care/harm domain is associated with the virtues of kindness and gentleness, and with vices such as cruelty. Norms about care and harm are familiar and ubiquitous. The simplest version of a care/harm norm is "Don't hurt innocent people."
2. Fairness/cheating: We tend to feel angry with cheaters and guilty when we ourselves cheat. Because our success as a species depended on our ability to cooperate with each other in small groups, it was adaptive for us to favor fairness and disfavor the cheating. The fairness/cheating domain is associated with the virtues of justice and trustworthiness and the vices of unfairness and dishonesty. Fairness norms are also very familiar: "Don't cheat," "Treat others fairly," and "Do your share of the work" are some examples.
3. Loyalty/betrayal: We tend to feel pride in our group membership and rage toward traitors. Again, due to our need for cooperation within our groups, loyalty to the group served us well. This domain is associated with the virtues of loyalty, patriotism, and self-sacrifice, and with the vices of infidelity and treason. Loyalty norms include "Put your family first" and "Defend your country."
4. Authority/subversion: We tend to respect (and sometimes fear) authority, which may have been an advantage when human beings started to live in larger, hierarchical groups. To make those groups work and to ensure continued cooperative activity, it helped to have a tendency to value doing what we're told. The authority/subversion domain is associated with the virtues of leadership, deference to legitimate authority and respect for traditions and to the vice of insubordination. Norms such as "respect your elders" are authority norms.
5. Sanctity/degradation: There is a human tendency to feel disgusted by rotten food and waste products, and also to deem certain people and behaviors sick or unhealthy. It's not hard to see how disgust at contaminated food was adaptive for hunter-gatherers, particularly as they began to eat meat, and at some point in our evolution, disgust reactions became moralized. The sanctity/degradation domain is associated with virtues such as temperance and chastity and the vice of indecency. "Treat your body like a temple" is an example of a loyalty norm, as is "do not debase yourself by using foul language" (adapted from Graham et al. 2013).

Before we consider these moral domains in more detail, a few caveats. First, as we've already seen, coming up with a plausible story about how a capacity may have been adaptive in evolution does not prove that the capacity is organized in advance of experience. An evolutionary explanation is not sufficient proof that something is unlearned. To show that these capacities are unlearned, Haidt and Graham would also want to show that they arrive early in development (prior to a bunch of experiential learning) for most humans across cultures. Second, many philosophers are highly critical of evolutionary explanations of our psychological capacities (Downes 2021; Dupre 2012; Lloyd 1999). In part this is because we do not have direct evidence about how these capacities evolved – typically, evolutionary psychologists get their evidence from psychological experiments on living people. This means that we might just be making up "just-so" stories rather than getting at a deep truth about unlearned human nature.

Let's put these caveats to the side for the moment in order to focus on the debate about what's in the rough draft of morality. Rottman and Young (whose research on childhood development we discussed above) endorse the taste bud metaphor (Graham et al. 2013) and they agree that certain moral capacities are unlearned:

> rather than needing to construct a moral sense through their own efforts, babies are born with certain prepared intuitions (the metaphorical taste receptors) that establish the boundaries for a mature moral sense. These intuitions are then modulated by cultural input, which adjusts the sensitivity of the receptors and the range of content to which they are responsive. This is therefore a form of nativism that allows for a large but finite degree of cultural variability …
>
> (p. 130)

But Rottman and Young don't necessarily agree with Graham and Haidt about *which* moral taste buds we have. Indeed, Graham and Haidt themselves think that their proposal about the first draft of the moral mind is itself a draft. They do not argue that their list is perfect – new domains may be discovered (they propose liberty/oppression as a likely candidate[7]) and future research may reveal that some of the domains on their list are not truly part of the rough draft but are more products of culture.

Moral Foundations Theory has been fruitful and influential, but – since even its authors think it is a work in progress – it is worth putting a viable alternative on the table. Victor Kumar and Richard Campbell have a proposal that is similar to MFT, but with important differences. We can call their theory "Gene-Culture Pluralism," because of the emphasis they put on the importance of genes *and* culture to how morality evolved. They argue that the five core moral norms are harm, kinship, reciprocity, autonomy, and fairness (2022: 93). You may notice some overlap here: harm, fairness, and kinship (similar to loyalty) are on both lists. There is also overlap on the moral emotions. Gene-Culture Pluralism includes the moral emotions of sympathy, loyalty, trust, respect, guilt, and resentment. Moral Foundations Theory also puts sympathy, loyalty, respect, and anger (similar to resentment) at the core of morality. You will also notice that purity and authority are absent from the core of morality according to Gene-Culture Pluralism.

So, there are important similarities between the two proposals discussed here, but there are also at least two significant differences. First, the two proposals have different views about innateness or what is unlearned. Graham and Haidt say that the moral domains, including the moral norms that characterize them, are innate; Kumar and Campbell explicitly reject this claim:

> There is no need to posit innate norms to explain what different moralities have in common... All that's needed are two innate capacities. One is a capacity to learn the norms found in one's culture: to quickly and easily gain intrinsic motivation both to follow them and to sanction violations. The other innate capacity is a set of resonant moral emotions.
>
> (2022: 94)

In other words, Graham and Haidt think that we are given *more* prior to learning than Kumar and Campbell think we are given. One way of thinking about the difference here is that it is a difference about the nature of the rough draft. Graham and Haidt take the draft to be a true rough draft, written in such a way that we are constrained to accept certain kinds of moral rules. For Kumar and Campbell, it's more like an outline that gives culture and learning a larger role in filling in the details of what specific norms our moral emotions support. Nevertheless, in both cases, we are prepared, prior to experience, to develop morality that has a certain shape. And in both cases, that shape is what it is because throughout our evolutionary history we had to coordinate with other people to take care of our helpless children, cooperate, and share in order to survive.

So, Moral Foundations Theory and Gene-Culture Pluralism make different claims about how detailed the draft is. They also differ over what the draft says, or which taste buds we have. As we've seen, Kumar and Campbell don't think that the purity and authority domains are a part of the core of morality. They think this, in part, because of the evolutionary evidence. Briefly, they argue that purity and authority norms did not become important in human evolution until we began to live in very large groups; they were not part of the morality of hunter-gatherer groups in early human evolution, and there is no counterpart to them in the rest of the animal kingdom. Disgust (the emotion associated with purity) was present very early, but (they argue) it concerned contaminated food, not people. Disgust had no moral significance in our early history. Kumar and Campbell also think that kinship and reciprocity deserve to be singled out as moral domains because of evolutionary evidence, whereas MFT lumps these in with loyalty and fairness. Developmental psychologists seem to be more on Kumar and Campbell's side of the debate, at least insofar as purity is concerned, because very young children do not, as far as we know, respond to purity and contamination in the ways they respond to benefit and harm; more research needs to be done here, but it seems that this aspect of morality is more likely to be learned (Aznar, Tenenbaum, & Russell 2022; Rottman & Young 2015; Stevenson et al. 2010).

As you can see, there isn't consensus about what the rough draft of morality looks like – and the two alternatives considered here are not the only possibilities (see O'Neill 2017 for an overview of options). That said, since there is very broad agreement about the care/harm domain, let's just recap what it means to say that this domain is "innate." It means that human beings are born prepared to care about harm and to sympathize with suffering in the right kind of environment. It does not mean that every human being will naturally develop into a utilitarian hero; it doesn't even mean that every human being will develop a sense of the wrongness of harm. Just as a child who never tastes salt will not know anything about salty food, a child who is raised by cruel sadists may not come to care about the suffering of others. But in most cultures and families with which we are familiar, children do easily sympathize with others. A child who grows up in a culture with no high fructose corn syrup and a child who grows up eating American junk food may have a different sense of what's a good amount of sweetness. Analogously, different expressions of sympathy might be encouraged in different cultures. The draft of morality is rough enough to accommodate a good bit of variation about what counts as harm and what one ought to do about it.

What Does It Mean for Ethics?

The science of moral human nature is not settled. My bet is that it will turn out that parts of morality are present in human beings prior to learning,[8] but I wouldn't take a bet right now on exactly which domains we have and what they look like. I don't think we have to know exactly what is in the rough draft, however, to ask what it means for moral philosophy that there is one. So, what does it mean? Do the starting points provide a foundation for morality? You might think so from the name Moral *Foundations* Theory. If there is a draft, how much can it be rewritten? Broadly stated, the question here is about the relationship between our moral nature (whatever that turns out to be) and normative ethics.

As I suggested in the introduction, there are three basic answers to this question: (1) the empirical facts (about unlearned moral capacities, in this case) have nothing to do with ethics, (2) the empirical facts by themselves ground or determine what is ethical, or (3) the empirical facts are relevant to ethics, but do not by themselves determine what is ethical. I continue to think that the first answer is implausible. How could it make no difference how we start out in life? At the very least, the fact that we are sympathetic creatures is relevant to the kind of morality that could possibly take hold for humans. Indeed, even David Hume – whose famous dictum "you cannot derive an ought from an is" motivates skepticism about grounding morality in science – thought that our natural sympathy is crucial to ethics.

If our moral nature is relevant to normative ethics, in what way is it relevant? One possibility – option 2 – runs afoul of Hume's is/ought dictum, but it's worth considering. There are some scientists and philosophers who think that the empirical facts determine what is ethical. In 1975, biologist E. O. Wilson said that "the time has come for ethics to be removed temporarily from the hands of the philosophers and biologicized" (Wilson 1975: 562). More recently, Sam Harris in *The Moral Landscape* (2011) argues that science will tell us what's morally right and wrong – indeed the claim that "science can determine human values" is in the subtitle of the book. The idea here seems to be that the empirical facts *settle* normative questions about what we ought to do, that the scientific method is the method for ethics.

Many of the philosophers and scientists we have considered in this chapter would reject this way of thinking. Why? To see why it's problematic, let's consider Moral Foundations Theory again. In *The Righteous Mind*, Haidt (2012) argues that the moral domains are distributed differently across the American political spectrum. The moral lives of political liberals tend to be more centrally focused on harm and fairness norms and the associated moral emotions of sympathy and anger. Conservatives, on the other hand, experience morality in a way that is more broadly distributed across all five domains. They are much more likely to endorse norms that prescribe respect for authority, loyalty, and purity.

In their review paper, Graham and Haidt are very clear that they are *describing* our morality rather than prescribing anything. Nevertheless, elsewhere Haidt (2012) does suggest that the descriptive picture he paints will provide grounds for a more positive appraisal of moral frameworks that are different from our own (that is, different from the perspectives of the western scientists and philosophers likely to be reading Haidt's work). Certainly, some American conservatives have taken him to be saying that the fact that they use more domains of morality is a *good* thing about political conservativism. As Haidt himself acknowledges in an interview about reactions to his book, "The reviews on the right say: 'Hey, conservatives, you should all read this book because it shows that we have more moral foundations than they do. Nah, nah, nah, nah, nah'" (Goldman 2012). But how do we get from the fact that some people rely on a broader range of sentiments in their moral judgment to the conclusion that such people are better?

Consider this argument for a moral conclusion:

1. All of Smith's moral judgments are based on harm or respect norms.
2. Jones's moral judgments are based on harm, respect, authority, or sanctity norms.
3. Jones relies on more norms to make moral judgments.

∴ Therefore, Jones makes better moral judgments than Smith.

As Hume would point out, there's a premise missing in this argument, because there is no judgment about what is better or worse in premises 1–3. What could this missing premise be? Here's a possibility:

4. More norms result in better judgments.

This would forge a connection between the premises and the conclusion, but is it true? I don't know. How would you go about showing that it is true? Premise 4 is an evaluative premise, so we have to assess it using the tools of moral philosophy: clear thinking, comparison across cases, and consideration of all the relevant facts and values. Taking this approach, we can ask: do more sentiments always improve our judgment? No. Sometimes anger can lead us to blame people for things they didn't do, fear can lead us to make mistakes about how much risk there really is, and love can blind us to the faults of our loved ones. So as a blanket statement, "those who rely on more sentiments make better judgments" doesn't seem correct. We need to think about which sentiments and in what context. We also need to think about what constitutes *better* moral judgment, which depends on how we conceive of what we are judging; what counts as better moral judgment depends on what you think the moral truths are that we are talking about when we make moral judgments.

Haidt does, ultimately, make a normative argument for the importance of the moral domains that conservatives use more than liberals. He argues that the sanctity foundation—the domain of morality associated with the sentiment of disgust—allows us to maintain a sense of what is sacred, which in turn "helps bind individuals into moral communities. When someone in a moral community desecrates one of the sacred pillars supporting the community [the kind of action that can evoke disgust, such as burning a national flag], the reaction is sure to be swift, emotional, collective and punitive" (Haidt 2012: 268). Haidt's argument does rely on a moral premise, namely, the premise that being bound to a moral community in this way is a good thing. This is a plausible premise, though we might ask whether it's enough to get to the conclusion. To get to the conclusion that disgust is a basis for sound, moral judgments, the argument also needs to assume that this disgust-based binding is worth whatever costs it may incur, such as exclusion of people who are deemed disgusting due to their sexual orientation, gender identity, class, or ethnicity.[9]

Kumar and Campbell also make an argument for (normative) ethical conclusions that is influenced by their (descriptive) scientific theory of the moral mind. The way they go about it is to state their evaluative premises as assumptions at the outset, and then to use the science of morality to inform how we can make progress. For them, moral progress means greater inclusivity and greater moral equality. We make moral progress when we expand the circle of those whom we take to be morally considerable, and also when we reduce relationships of domination and subordination within this circle. Kumar and Campbell use examples such as the end of chattel slavery in the United States, and the expansion of suffrage, to argue that these evaluative judgments "are about the sturdiest possible" (2022: 189).

With the ideal of moral progress in hand, they argue that *rational moral change* is the best route to making progress, given our innate capacities to feel moral emotions and to learn, follow, and enforce the norms of our culture.

Rational moral change involves adjusting our moral norms on the basis of a consistent application of accurate beliefs about the world:

> [W]hen people form accurate beliefs about the world around them and those who inhabit it, they tend to re-evaluate their moral feelings and norms in ways that lead them rationally toward greater inclusivity and equality. For example, dehumanizing and subordinating ideologies rest on factual mistakes about the people they exclude or demand, often compounded by an inconsistent application of shared core moral norms.
>
> (pp. 195–196)

Kumar and Campbell make a good case for this optimistic position, but you don't have to accept their conclusion to see the wisdom in the type of argument they are making. The nature of the moral mind will shape what kind of change it makes sense to hope for, and what methods it makes most sense to use to get there. The discovery that we are naturally inclined to care about other people and that we are capable of changing the specific norms we hold is important to arguments about how to make moral progress. If we were stuck with limited capacities, or if our minds were less flexible, rational moral progress may not even be possible. That is certainly true. It is also true, as Kumar and Campbell acknowledge, that we do not establish the *value* of inclusiveness or equality by locating them in the mind.

Whatever you think of these particular arguments, the point is that in order to conclude that it's good to have more moral taste buds, or to take steps to become more morally inclusive, we need some premises that are not purely a description of how things are. If there are moral lessons to be drawn from science, the arguments that support these lessons will have to include both scientific and normative premises. If you agree, then (like me) you accept the third answer to the question about science and ethics: empirical facts are relevant to ethics, but do not by themselves determine what is ethical. The word "relevant" in this answer is awfully broad. There are all sorts of ways that science might be relevant to ethics. One way we have just discussed is that knowledge of the rough draft can inform us about how best to work on revisions. In the rest of this book we will see many more examples.

Taking Stock

At the beginning of this chapter, I suggested that there are two philosophical traditions that take human nature to be the key to ethics. According to one, our nature defines what is ethical. Against this view, I've argued that we can't infer evaluative conclusions directly from empirical premises alone. Moreover, we've seen that even if the nativists are correct that morality is built into our nature, whatever is built in is so flexible that it wouldn't give us much specific guidance anyway. That said, it is useful to know what we start with. The fact that we are predisposed to have sympathy and to care about each other, for example, is surely relevant to what kind of ethics we should have, even if it doesn't dictate the exact shape of morality. According to the second tradition, morality is the enforcement mechanism we need to curb our selfish nature. From the evidence in this chapter, that doesn't really seem true either. We're not done with selfishness, however. In the next chapter we'll consider a different kind of argument for this "egoistic" tradition in philosophy, which will require us to look at a different body of empirical evidence.

Summary

- To answer the question whether morality is innate, we need definitions of "morality" and "innate."
- One way to define "morality," favored by psychologists featured in this chapter, is by observing the ordinary practices we tend to think of as moral.
- "Innateness" is a concept that tends to connote inflexibility and hard-wiring. Whatever moral capacities we are born with, they are not hard-wired. For this reason, we should either avoid the word "innate" or define it carefully.
- Moral nativism is the view that there are unlearned moral capacities that will tend to develop in ordinary environments. Moral domains, taste buds, or elements of the "rough draft" channel moral development in certain directions, but these channels may be wide enough to leave a lot of room for variability depending on a person's culture and experience.
- One kind of evidence for unlearned moral capacities is from the early emergence (in babies as young as 6 months) of preferences for helping behavior and against hindering behavior.
- Another kind of evidence for unlearned moral capacities is from evolution. Morality seems to have evolved to enable us to reap the benefits of cooperation.
- There are different theories of what moral domains we evolved to have. Moral Foundations Theory posits five domains (care, fairness, loyalty, authority, and purity). Gene-Cultural Pluralism lists harm, kinship, reciprocity, autonomy, and fairness. These two theories are similar, but they have different views about which moral domains we have and the extent to which these domains are unlearned.
- Whatever the details of the starting points of human morality, it is a good bet that we do not start with a blank slate and that culture and learning have a profound influence on how those starting points develop.
- The moral starting points are relevant to ethics, because what we are like at the very least constrains what we ought to do and what we can aspire to be. However, we cannot draw conclusions about what is right and wrong only from the facts about what moral taste buds we have. Normative premises are needed to establish normative conclusions.

Study Questions

1. What is at stake in the debate about whether morality is native or learned? Does the resolution of this debate matter to how we should think about our moral obligations or moral progress?
2. Think about your interactions with young children (if you have had some). What sorts of moral capacities have you observed and what features of moral agency are they missing?
3. Do you think our feelings of disgust tell us anything morally significant? If not, why not? If so, what do we do about the fact that disgust can veer into prejudice?
4. How might you design an experiment to test whether the moral domains of authority or purity are found in young children?
5. What kinds of scientific discoveries would make you more or less hopeful about our prospects for moral progress?

Notes

[1] *This American Life*, "No Fair," episode 672. Last accessed June 21, 2022: https://www.thisamericanlife.org/672/transcript

[2] In Western philosophy, Aristotle and Hobbes are the most obvious representatives of the two sides; in Chinese philosophy, it's Mencius and Xunzi.

[3] See also, Griffiths' (2002) classic critique of "innateness."

[4] Looking time is also a method used by developmental psychologists. Sometimes looking time is used as a proxy for preference – babies are thought to look longer at what they like than what they dislike. In other experiments, looking time is used to draw inferences about what babies expect to happen, the idea being that babies look longer at surprising or unexpected events. For a thorough discussion of the pros and cons of looking time research see Aslin 2007.

[5] For a review, see Woo et al. 2022; for alternatives to nativism see Tasimi 2020; Rhodes & Wellman 2017. Not everyone agrees that these studies have been sufficiently confirmed; see Schlingloff, Csibra, and Tatone (2020) on the controversy.

[6] There is a variety of moral nativism that puts less emphasis on social learning than the experts we have been talking about. These are nativists who rely on an analogy to linguistic ability and who argue that there is a "universal moral grammar." See Mikhail 2007 if you're interested in learning more about this option, and for a critical perspective see Sterelny 2010.

[7] The liberty/oppression foundation concerns feelings of resentment toward people who restrict our freedom and norms against bullying and tyranny.

[8] As you might expect in philosophy, not everyone agrees. Jesse Prinz (2009) is a staunch critic of moral nativism; he thinks morality is entirely learned from culture.

[9] Philosophers such as Dan Kelly (2011) and Martha Nussbaum (2009) have argued that disgust leads us morally astray.

Further Reading

Dahl, A. 2019. The science of early moral development: On defining, constructing, and studying morality from birth. *Advances in Child Development and Behavior 56*: 1–35.

Graham, J., J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, & P. H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology* (Vol. 47), pp. 55–130. Academic Press.

Hamlin, J. K., K. Wynn, & P. Bloom. 2007. Social evaluation by preverbal infants. *Nature 450*(7169): 557–559.

Kumar, V., & R. Campbell. 2022. *A Better Ape: The Evolution of the Moral Mind and How It Made Us Human*. Oxford University Press.

Rottman, J., & L. Young. 2015. *The Moral Brain: A Multidisciplinary Perspective*, pp. 123–142. MIT Press.

Tomasello, M. 2016. A natural history of human morality. In *A Natural History of Human Morality*. Harvard University Press.

3 Moral Motivation and Selfishness

In the last chapter we learned that human nature probably isn't all bad. But it's a long way from this claim to the claim that people are genuinely moral beings, capable of real moral motivation. For one thing, we haven't really talked about what real moral motivation is yet! So far, we have put on hold questions about the content of morality and the precise nature of moral motivation. (In the last chapter we just went with the "you know it when you see it" approach to defining morality.) As we also saw in the last chapter, to think about what's morally required, we need to do some moral philosophy – we don't find out which motivations are the *right* ones by empirical investigation alone. As we delve into these topics, we'll see that there is an empirical challenge to the claims some moral theories make about motivation, one which highlights our self-interested nature.

Moral Theories and Moral Motivation

Harry and Susan have something very rare in common: both have donated a kidney to a total stranger. Here's Susan's explanation of her action: "I believe I should try to help people. This seems to be a perfect opportunity to help someone in a big way with minimal inconvenience to myself" (McCann 2012). And, yes, she does know that "minimal inconvenience" includes a 1 in 2,500 risk of death during surgery. Harry's most frequent response when he is asked why he did this is to say that "it was the right thing to do" (Steinberg 2003).

To understand moral motivation, we need to know what a moral action is. Altruistic donation provides a fairly clear case, because it is a case in which a

DOI: 10.4324/9781003402787-4

person acts in a way that benefits others at some cost to herself for what look like morally good reasons. These are hallmarks of moral action. But are these necessary features of moral action? Does all moral action benefit others? Do all moral actions involve self-sacrifice? Is an action only moral if it's done for the right reasons? Not necessarily. According to some moral theories, moral action is not the same as action that produces beneficial consequences. According to some moral theories, the fact that an action is in your self-interest does not disqualify it as a moral action. And according to some moral theories, acting for the right reasons is not necessary for acting morally.

For our purposes, we can divide moral theories into four types, each of which has distinct implications for what counts as moral motivation. Let's start with utilitarianism. According to utilitarianism, a consequentialist theory, the right action is the one that produces the greatest welfare or happiness for the greatest number in the long run. Typically, utilitarians have not thought that it matters very much what motivates someone to produce happiness; producing the best consequences is the right thing to do no matter what your reasons for doing it. There are certainly consequentialists who care about motives, but they identify good motives in terms of their reliability at bringing about the best consequences (Driver 2001).[1]

Kantian moral theory takes quite a different position on the importance of motives. According to Kant, the only actions that are morally worthy are actions done from the motive of duty. The motive of duty is the motive of a good will, which is the kind of thing that is always good to have. It is, Kant thinks, unconditionally good:

> A good will is not good because of its effects or accomplishments, and not because of its adequacy to achieve any proposed end: it is good only by virtue of its willing – that is, it is good in itself … . [I]f with even its utmost effort it still accomplished nothing so that only good will itself remained (not, of course, as a mere wish, but as the summoning of every means in our power), even then it would still, like a jewel, glisten in its own right, as something that has its full worth in itself.
>
> (Kant 2002/1785: 394/196)

When we act from duty, we act with the intention – not just an inclination, but the "summoning of every means" – to do the right thing, because it is the right thing, no matter what we feel like doing. A person who acts from the motive of duty does the right thing in virtue of her recognition that morality demands it and her background rational commitment to do what is morally required of her. The motive of duty is important not only because it is the only morally good motive, but also because morality only applies to beings who are capable of being motivated by duty (that is, according to Kant, rational beings).

Kant's position may sound rather extreme, but the insight that a morally admirable motive is one that will get you to do the right thing no matter how you happen to be feeling is a good one.[2] To see this, think of someone who gives money to charity simply because she feels sympathy for the victims of a natural disaster. It's not that she believes this is the right thing to do, nor that she is concerned to do the right thing – she is just, say, feeling sorry for the people she saw on television and responding from her gut. Certainly she isn't doing anything wrong, but is what she's doing admirable? Consider that her sympathetic motive might be quite fickle. When the television stops showing the suffering, she stops helping; when suffering people appear on TV who do not engage her sympathies (perhaps because they are from a different continent or follow a different religion from her), she doesn't help; when her sympathy is swamped by feeling annoyed at her boyfriend for forgetting her birthday, she doesn't help. The motive of duty is supposed to be much more dependable and consistent than this. The motive of duty can motivate us even when the going gets tough; because it is independent of our feelings and desires, duty can motivate us to do the right thing even when we're feeling lazy, or mean, or we want to do something else. Intuitively, this sounds like a motive that's worth praising.

Motives are important for Kantians. It can look like we never do anything right at all unless we act purely from the motive of moral duty. But this is probably not what Kant meant. There is room, in Kantian moral theory, for a distinction between right action and morally worthy action. The right action is one that conforms to the supreme moral principle, the categorical imperative, which tells us to treat others as unconditionally valuable ends in themselves, never merely as means to our ends. Morally worthy actions are right actions done from the motive of duty. On this interpretation, we can do the right thing from an unworthy motive. For example, if I keep my promise to meet you for dinner, I have treated you in the way you deserve as a rational being, and, according to the categorical imperative (the supreme moral principle), I have done the morally right thing. But I may not have been motivated by duty; instead, I might have kept my promise because I wanted to eat at that restaurant anyway. In this case, I will have done the right thing, but not from morally admirable or worthy motives. So, the Kantian is not stuck saying that in keeping my promise for the wrong reasons I did the wrong thing. Still, Kantian moral theory enjoins us to do the right thing *for the right reasons*, that is, from the motive of duty. The important point for our purposes is that, according to the Kantian view about moral motivation (where this means motivation that is worthy of admiration and not just whatever happens to get us to do the right thing), it must be possible for us to act independently of our inclinations. Notice that this makes the Kantian picture depend on a claim about human psychology, namely, that we are capable of acting against our inclinations. Whether this claim is tenable and just how much it matters for Kantianism will be considered in detail in Part II of this book.

According to a third type of moral theory, contract theory, morally right action is defined in terms of a hypothetical contract between moral agents. There are many different versions of contract theory, but the one I want to focus on here is T. M. Scanlon's contractualism. According to Scanlon (1998), the right thing to do is the action that follows the principles that no one could reasonably reject as a basis for regulating our interactions with each other. In Scanlon's view a right action could be done from the wrong motive: if your action is in accordance with the principles, then it's morally right regardless of your motive. But Scanlon thinks that any moral theory must explain why we should care about the imperatives that it offers; therefore, he pairs his theory with a view about moral motivation that fits together with it naturally. Scanlon believes that we desire to justify ourselves to others on grounds they could not reasonably reject and that this desire explains how we feel about certain aspects of morality. For example, Scanlon (1982) reports that his own response to not meeting his obligation to help suffering people in distant places (and what might motivate him do more) is not to feel bad out of genuine sympathy for those other people, but to feel like he couldn't really justify his lack of attention to the suffering people given how easy it would be for him to give more. In his experience, this desire for his actions to be justified to others is an important moral motive. Contractualism does not hold that this desire to justify ourselves to others is the only moral motive, but it is an important one because it provides a reason for us to care about the hypothetical contract. Thinking back to the previous chapter, this desire to justify ourselves to others is also likely to be a learned development from our natural sympathy and sense of fairness.

Finally, virtue ethics takes moral motivation to be extremely important because being motivated in the right way is essential to being a virtuous person who lives a flourishing life. (Much more will be said about virtue ethics in the next chapter and in Chapter 9). According to the most prevalent version of virtue ethics, the one influenced by Aristotle, the best life for a human being is the life of virtue, because virtues (such as courage, temperance, justice, and wisdom) are the excellent exercise of our human capacities and the best life for human beings is the life spent exercising these capacities in the most excellent way. Importantly for our purposes, virtues are partly constituted by certain kinds of motivating states, like desires and emotions. For example, a truly generous person wants to give people things and feels pleasure in doing so, which motivates her to do it again. Some virtues are defined in terms of the right amount of emotion: a courageous person, for example, feels fear, but not too much or too little and not in the wrong circumstances. Moral motivation, according to virtue ethics, is virtuous motivation; it therefore requires desires and emotional dispositions that are appropriately tuned and governed by wisdom.

Different moral theories have different views about what kinds of motivation are good or praiseworthy, as well as about whether motivation matters to getting the action right in the first place. According to Kantianism, an action isn't a morally worthy action if it is done for the wrong reasons, and according to virtue ethics, right action is action that stems from virtuous motives. Contractualism assumes that we are motivated by a desire to justify ourselves to others and, while it is possible on this view to act rightly without being directly motivated by this desire, the theory is only a plausible moral theory (according to one of its main proponents) if we do indeed have this desire. In subsequent chapters we will want to see if these claims about our motivations are well founded. If they are not, we have some reason to abandon the theories that make them. It is surely a count against a moral theory if it makes implausible assumptions about moral motivation.

Now, you might think that we could make things easier for ourselves if we stick with utilitarianism and investigate why people sometimes promote general happiness and sometimes do not. This does seem to be the starting assumption of a good deal of psychological research about moral motivation; much attention in psychology is devoted to investigating the causes of "pro-social behavior," which is behavior that produces good outcomes understood in basically utilitarian terms. If our sole interest were in promoting pro-social behavior or figuring out how to make people more helpful, this wouldn't be a bad strategy. In other words, if our interest in moral motivation were itself consequentialist (if we just want to know what motivates people to act in a certain way so that we can produce more of that kind of action), then assuming that consequentialism is the right moral theory makes sense. But, if consequentialism doesn't explain everything there is to explain about moral action, then we would miss something by assuming that it is true.

More importantly, even if motives don't determine which actions are morally right, many people care deeply about motives and it is worth thinking about which motives are relevant to the question of how to live. That is, once we distinguish rightness of actions from moral motivation, we can see that both are important. To see this, consider some examples. Picture a philosopher who goes to visit a very sick friend in the hospital, and the friend asks why she has come. The philosopher tells her friend that she thought perhaps her presence would remind him to pay back the five dollars he owes her. Or imagine the philosopher is on a sailboat with her boyfriend who falls overboard. She later reports that she saved him rather than the drowning stranger next to him because she had no way of knowing which of them would contribute more during their lives to the greatest happiness and, therefore, the imperative to maximize expected happiness permitted her to save her own boyfriend.[3] In both cases, the philosopher did the right thing, but we probably wouldn't want her for a friend. There is something wrong with people who have these motives, rather than the motives of friendship, sympathy, and love.

If we think about the kind of friends we want to have, or the kind of friend we want to be, *how* we are motivated to help seems just as important as whether we end up helping. A gift given by a friend out of the expectation of a gift in return is not as nice as a gift that your friend gives you because she thinks it will make you happy. A person who gives to others freely and with pleasure seems more admirable than one who gives grudgingly. And most people would rather raise children who are sympathetic, kind people who have good friends and at least enough of a sense of duty to keep their promises and vote in elections. I hope these cases are enough to motivate the thought that we need to look beyond the causes of pro-social behavior in order to understand moral motivation in its fullest sense.

One thing about all of these motives we tend to like and admire is that they are not selfish, and this might set off some alarms. You might be thinking that all this deliberation about motives – which ones are praiseworthy, which ones we should seek in our friends, which ones we should cultivate in ourselves – is irrelevant because we are at the core selfish creatures who can only be motivated by the desire for our own good. If we are always selfish, then most of the theories we've discussed so far make false assumptions about how we can be motivated. If I always act for the sake of number one (me), then I cannot act for the sake of duty, or from a desire for someone else's good, or from a desire to justify my actions to others whose regard I esteem, or out of genuine virtuous compassion. It is therefore worth considering whether there is some truth to the idea that we are always selfish. This is the challenge we will consider in the remainder of the chapter.

The Challenge of Psychological Egoism

Psychological egoism (henceforth just *egoism*) is the view that all voluntary action is motivated by self-interest, or, to put it another way: we always act selfishly. The opposing view, I will call *non-egoism*, is the view that *not all* voluntary actions are motivated by self-interest. You can see why egoism challenges some of the moral theories which we just considered. If we always act selfishly, then we do not ever act from a sense of duty, because duty motivates us independently of all our desires and inclinations and, therefore, independently of our selfish desires. If we always act for the sake of our own interests, then there are certain other-regarding virtues that we cannot possess; for example, we cannot help another person for her own sake, as seems to be required by benevolence.

Sometimes psychological egoism has been put forward as a conceptual truth. Thomas Hobbes (the English philosopher whose defense of absolute government authority in *Leviathan* was published in 1651) is sometimes taken to have made this argument. Hobbes tells us that the object of every person's voluntary act is "some good to himself" and this seems to follow from his definitions of "good" and "voluntary action." A voluntary action is one that proceeds from the will, which is itself the last appetite or aversion before the action occurs. The good is just the object of our desires: "whatsoever is the object of any man's appetite or desire that is it which he for his part calleth *good*."[4] According to this interpretation of Hobbes, when we act voluntarily, we act on a desire for what we take to be our own good. To say that egoism is a *conceptual* truth is to say that it follows from the meaning of "voluntary action," "desire," and "good." We do not need to establish egoism empirically; we already know that it must be true.

The idea that we know egoism is true without having to provide evidence for it underlies a pattern of argument with which many readers may be familiar. It goes this way:

Egoist: We always act selfishly.

Non-Egoist: No, we don't! Look at Mother Teresa!

E: Mother Teresa was just trying to get into heaven.

NE: Maybe, but what about the soldier who falls on a grenade to save his platoon?

E: He did it to avoid feeling guilty if he didn't.

NE: What about the time I helped a friend move even though I really didn't want to?

E: You are deceived about your own motives. You obviously really did want to, or you wouldn't have done it! Probably your real motive was to get the pleasure of feeling like a good person for helping.

Things could go on this way for some time until one person (typically the non-egoist) ends up too frustrated to proceed. What we should notice about this pattern of argument is that, without acknowledging it, each side is putting forward claims about the explanations of actions that are supposed to be matters of fact, but neither side is offering any evidence for what is supposed to be a fact (with the possible exception of the non-egoist's last example in which she offers evidence from introspection). The egoist claims that Mother Teresa helped the lepers in order to get into heaven. How does the egoist know this? What kind of evidence does the egoist offer to support this idea? Typically, no evidence is provided, because it seems obvious to the egoist: it *has to be* that Mother Teresa had some ulterior selfish motive, because otherwise why would she have done it? This seems obvious to the egoist – and empirical evidence seems unnecessary – because underneath the factual claim about what explains Mother Teresa's action is a conceptual claim that the egoist believes, which is that the only way an action could possibly be produced is by way of a selfish desire.

Why does this seem obvious to the egoist? Why does it seem like actions *must* be selfish and that this is something we can know from the armchair? The answer is that there is a premise in these arguments that is at least plausibly a conceptual truth. It's just that this premise does not support egoism by itself. The premise is that "all voluntary actions are caused by desires." We can see this premise at work in the above dialogue when the egoist says, "You obviously really did want to, or you wouldn't have done it!" This could be true, but it doesn't follow that your action was selfish unless we add to the argument the further claim that what you wanted from your action was your own pleasure (or something else for yourself). The (conceivably conceptually true) premise here is that desires are necessary to motivate actions, but what is incorrectly inferred from this premise by the egoist is that only *selfish* desires motivate action. That's the mistake. The idea that desires are necessary to motivate action is a very reasonable idea, and it's one that we will consider in some detail in Chapter 5. The important point for our purposes in this chapter is that it does not follow from the fact that we act on our desires that we act selfishly. Again, this is because it doesn't follow from the fact that desires cause action that *selfish* desires cause action. The missing premise – a controversial premise for which the armchair egoist provides no evidence – is that all desires are selfish.

Psychological Egoism and Empirical Research

So, if egoism is true, then it is an empirical fact, not a conceptual truth. Is it true empirically? How would this be investigated? To put it in the same terms as the argument we discussed above, what we need to ask is whether desires are always desires for something for oneself. At first glance, this seems obviously false. We want all sorts of things that aren't for things for ourselves. People want their friends to be happy and their enemies to suffer. We want our parents to be healthy and our spouses to enjoy their jobs. We want world peace and a solution to climate change. In none of these cases is the immediate object of our desire a benefit for us. It does seem like we observe myriad examples of desires that are not selfish.

Certainly there are unselfish desires, but it's not clear how deep this point is. Here we need to make a distinction between instrumental desires and ultimate desires. An instrumental desire for x is a desire that depends on a further desire for something else to which x is a means. An ultimate desire is a desire for something for its own sake, not because it is a means to anything else. For example, the desire for money is (for most people) an instrumental desire: we want it for the sake of the things that money can buy, not for itself. A desire for one's own happiness, on the other hand, is an ultimate desire: we don't want to be happy because it promotes any other goal we have, we just want it for itself. Now we can ask about the examples that confront us whether they are examples of instrumental or ultimate desire. Do we want our parents' health for its own sake (or for their own sakes?), or do we want it because it makes us happy or reduces the burden they are to us? Do we want world peace for its own sake, or do we want it because it would make our own lives more pleasant? It is this question about *ultimate* desires that needs to be answered.

You might think we could find out about people's ultimate desires simply by asking them what they really want for its own sake. But people are notoriously bad at knowing exactly what they want ultimately or "deep down," and, as we saw above, the egoist is likely to suspect that those who claim to have an altruistic desire are deceived about what they really want. So, we should try to find some other methods. One way to try to answer this question about ultimate desires would be to ask how likely it is that human beings have altruistic ultimate desires (desires for the well-being of others) given what we know about human evolution. As we discussed in the previous chapter, there are plausible explanations for how we might have evolved to have non-egoistic, ultimate desires such as a desire for our children to thrive. Elliott Sober (a philosopher of biology) and David Sloan Wilson (an evolutionary biologist) argue that other-regarding desires could have been selected for by evolution because they are a more reliable way to get us to care for our children than egoism.

Now you might think that desires for our children's well-being aren't really altruistic because they contribute directly to the passing along of our own genes. But this would confuse two very different kinds of explanation. On the one hand, human beings might have evolved (and did, according to Sober and Wilson) to care about their children's well-being because doing so made it more likely that their genetic material would be passed on to future generations. We explain why human beings have altruistic desires by appeal to evidence about which characteristics are adaptive. On the other hand, the psychological explanation for why a particular person takes care of his children might be (and is, according to Sober and Wilson) that he wants his children to fare well, for their sake, completely independently of any concerns he might have to reproduce or pass along his genetic material. The explanation for why human beings are the way we are refers to which psychological states benefitted people in terms of survival and reproductive success. But the psychological states that evolved may not themselves make any reference to survival or reproduction.

To see this, think about our aversion to pain. It seems likely that we evolved to dislike pain – and to want to avoid it – because pain is usually a signal that we have sustained damage that might affect our survival and chances to reproduce. But the desire not to be in pain does not itself have anything to do with surviving to reproduce, that is, the concern to survive and reproduce isn't built into the psychological state of wanting to avoid pain. When you're in pain, you're not thinking about your survival and reproductive options; you're thinking, "Ow! That hurts! Make it stop!" To put the point another way, if you are asked to explain your desire to avoid pain from your own point of view (not from the point of view of the evolution of the human species), you will probably be nonplussed. But if you really had to try to explain it, you would probably say something like "pain is awful" or (even more simply) "pain sucks." The fact that our psychology is what it is because of evolutionary pressures does not make those evolutionary pressures any part of what the psychological states that we experience are *about* (evolution isn't part of the content of these mental states).

If Sober and Wilson's argument works, then, we do have reason to think that human beings evolved to have altruistic desires.[5] And, once we recognize the point that the contents of our psychological states (what they are about) are not the same as the evolutionary pressures that created them, if we can have altruistic desires for our children's well-being, then there's no reason why we couldn't have them for other people as well. Does Sober and Wilson's argument work? It has certainly been criticized (Stich 2007; Stich, Doris, & Roedder 2010). One serious problem is that while the evolutionary argument may show that altruistic desires are compatible with evolution, it doesn't provide direct evidence of altruistic desires. For that we have to look to social psychology.

We do find evidence for the existence of altruistic, ultimate desires in social psychology. Daniel Batson and his colleagues have developed a series of ingenious experiments to test what they call the "empathy-altruism hypothesis," according to which empathy – an "other-oriented emotional reaction to seeing someone suffer" – leads to altruistic action (Batson 1991). Empathy, Batson and colleagues discovered, can be induced in a person by getting her to take someone else's perspective or, in other words, by getting someone to imagine how another person is affected by their situation. And, further, empathy has been shown to increase helping behavior. The fact that empathy increases helping behavior does not amount to evidence against egoism. After all, it could be that empathy produces helping behavior by, for example, causing a desire to get rid of the pain one feels due to the empathy with someone else's pain. The non-egoist needs to show that the hypothesis that empathy increases helping behavior because it gives rise to an ultimate desire for another person's well-being is a more plausible hypothesis than the egoistic hypothesis. This is exactly what Batson and his colleagues aim to show.

We can see Batson's research as an attempt to move beyond the impasse between the egoist and the non-egoist that we saw in the previous section. The egoist has one hypothesis about helping behavior, the non-egoist has another, and neither has any evidence that convinces the other. Batson's approach is to test his non-egoistic hypothesis – the empathy-altruism hypothesis – against as many plausible egoistic hypotheses as he can. Since these experiments are rather complicated, we'll just discuss a couple of them here. Interested readers can look to the suggested readings at the end of the chapter to find out more.

One thought the egoist might have is the one I just mentioned: that empathy increases helping behavior because people want to eliminate the distress that empathy causes them. (If empathy makes us feel others' pain, then it is itself somewhat painful to experience.) On this view, which Batson calls the "aversive-arousal reduction hypothesis," we don't have ultimate desires for other people's well-being. Rather, we desire other people's well-being only insofar as someone else's well-being will produce more well-being (less distress) for us. To test this hypothesis against the empathy-altruism hypothesis, Batson set up a series of experiments that allow the subject to escape the scene rather than help. The thought is that if all the person desires is her own well-being, she would take the most efficient means to this end, which would be to escape from the suffering other (rather than to help her).

In these experiments, participants watch Elaine attempting to perform some tasks under "aversive conditions" (Batson et al. 1981). More specifically, they are watching Elaine suffer electric shocks (and, in fact, they are watching a video of someone pretending to suffer electric shocks) for the sake of an experiment on learning under aversive conditions. The subjects are told that Elaine is particularly sensitive to these shocks because of an early childhood trauma and that they could take Elaine's place if they wanted to. Half the subjects are primed to be empathetic with Elaine, half are not.[6] Half of each group (primed and unprimed) is presented with an obstacle to escaping (they are told that, if they do not take Elaine's place, they will have to watch a bunch more of these aversive condition trials); the other halves of each group are not presented with this obstacle. Given these variables, the participants are divided into four different groups:

1. Low-empathy, easy escape.
2. High-empathy, easy escape.

3. Low-empathy, difficult escape.
4. High-empathy, difficult escape.

The most important comparison for testing Batson's empathy-altruism hypothesis against the aversive-arousal reduction hypothesis is the comparison between groups 2 and 4. This is because the empathy-altruism hypothesis would predict that an easy out makes no difference to helping behavior of empathetic people because the subjects' ultimate desire is for Elaine to feel better, not for their own distress to be reduced. On this hypothesis, people in group 2 should be just as likely to help as people in group 4. On the other hand, the aversive-arousal reduction hypothesis predicts that people in group 2 will be more likely to escape (not help) than people in group 4, because escaping allows them to avoid the aversive feeling of empathizing with someone who is in pain. As it turned out, Batson's prediction was right: people in group 2 were at least as likely to help as people in group 4 (i.e., empathetic people helped Elaine whether or not they could escape easily). This is some evidence that empathy is functioning to produce an ultimate desire for Elaine's well-being rather than activating an ultimate desire for the subject's own well-being.

There is one obvious flaw with these experiments. Batson concludes that empathetic subjects in the experiment choose to help rather than to escape even when it's easy to escape because they have an ultimate desire to help Elaine. But this conclusion depends on the assumption that the subjects believe they won't go on thinking about Elaine and feeling distress after they escape the experiment. A person who thought he would continue to feel distress about Elaine after escaping may still be motivated by his ultimate desire for his own well-being (or lack of distress). Fortunately, another group of psychologists has done some experiments that support Batson's assumption (Stocks, Lishner, & Decker 2009). In these experiments, psychologist Eric Stocks and colleagues put people in similar situations to the ones that are used in Batson's experiments: they learn about a suffering person, Katie, with whom they empathize more or less (depending on which priming group they are in), and they can choose either to help Katie at some cost to themselves or to escape. Stocks adds a new variable to the experiment, which is that half the subjects are told that they are in a "deleting memories" training session that will make them forget everything they hear about Katie, whereas the other half are told they are in a "saving memories" training session and therefore they will remember what they learn.[7] If people were only motivated to help Katie in order to avoid personal distress, you would think that the fact that they aren't going to remember the distressing story would undercut their motivation to help. But this is not what happened in the experiment. Instead, the pattern Batson found was repeated: subjects who empathized with Katie were just as likely (in fact, a little more likely) to help if they thought their memories would be deleted as they were if they thought they would remember Katie's suffering.

Batson's empathy-altruism hypothesis, then, looks more likely than the aversive-arousal hypothesis (and the other hypotheses against which he tests it).

However, there is at least one other possible egoistic explanation of seemingly altruistic behavior that hasn't been much discussed. According to the "Big God" hypothesis, people act in ways that help others, because they are concerned about punishment from an omniscient and moral God (Shariff & Norenzayan 2007; Shariff et al. 2016). This would be an egoistic form of motivation that causes people to act morally to save their own skin. It's also worth noting that there are other *non*-egoistic explanations that compete with the empathy-altruism hypothesis. Principlism is the view that some actions are motivated by an ultimate desire to uphold a principle, such as a moral principle of justice (Batson 2011). Principlism is not egoistic, but it is an alternative to altruism as an explanation for moral motivation. Notice that if principlism were true, this would not cause the same problems for moral theories that egoism presents. Indeed, the existence of motivation by principle would support the Kantian theory of moral motivation very nicely.

If we listen to the main critic of Batson's research in philosophy, Stephen Stich, we will think that Batson's research has advanced the debate, though we do not yet have conclusive proof against egoism. Notice that *conclusive* proof will be very difficult to acquire, since it requires excellent evidence against every alternative egoistic hypothesis and, as Stich, Doris, and Roedder (2010) point out, evidence against all the possible *combinations* of egoistic hypotheses. Nevertheless, the empathy-altruism hypothesis does look plausible and we have no overwhelming reason to reject it given what we know now.

Taking Stock

Do we have genuinely other-regarding motivations, or do we always act selfishly? This is an important question for moral philosophy, because if we always act selfishly, some moral theories need to revise their views about moral motivation. It is possible that we always act selfishly, but if this is true it cannot be proved from the armchair. What kinds of motivations we actually have is an empirical question and scientific evidence is relevant to answering it. We have also discovered that the current science doesn't provide an easy answer to this question. Non-egoism is on pretty solid ground, though. There is a good bit of evidence that altruistic motivations plausibly evolved and that people do sometimes act on these motivations. Further, there isn't any evidence that contradicts the assumption that we have some non-selfish motives.

Notice that the best science available provides evidence for a particular hypothesis about the content of our non-egoistic motivations: that we have ultimate desires for the well-being of other people. As noted above, this research does nothing to prove or disprove hypotheses about other non-egoistic motivations, such as acting on principle (priniciplism) or the motive of duty. Nevertheless, if it is true that we have ultimate desires for others' well-being, many moral theories have reason to celebrate. Utilitarianism, while it does not build assumptions about motives into its definition of right action, should welcome the idea that we can be motivated directly by concerns for other people's happiness. Virtue ethics, because it prizes other-regarding character traits such as benevolence and kindness, will be on firmer ground if Batson's hypothesis is true. Contractualists, in general, believe that we have moral reasons to safeguard the well-being of others, so they should be delighted that we desire to do so. Even Kantians should be glad. After all, Kantians think that we have an imperfect duty to promote other people's happiness and the ultimate desire for the well-being of other people – while not a morally *worthy* motive – will at least help us to perform this duty.

Summary

- Different moral theories make different assumptions about moral motivation.
- From a traditional utilitarian perspective, motives do not matter to what we ought to do. We ought to maximize utility, and moral motivation is whatever gets us to do that.
- According to other moral theories, such as Kantianism and virtue ethics, motivation is crucial to what we morally ought to do.
- Psychological egoism (PE) is the view that *all* of our actions are selfish or motivated by self-regarding desires.
- Most moral theories assume that PE is false.
- It is a mistake to take PE to be a conceptual truth; PE is an empirical claim.
- The empirical evidence does not establish PE. The empirical evidence also does not prove conclusively that PE is false; however, there is a good deal of evidence that altruistic motivation would have evolved and that people sometimes act from altruistic motives such as the desire for another person's well-being.

Study Questions

1. What assumptions do moral theories make about human motivation? What are some examples of the different roles that these assumptions might play in a moral theory?
2. Think of an example of someone you admire doing something morally good. What do you think their motives were? Do you care? Are their motives part of what is admirable about them?
3. Sometimes in moral psychology, progress is made by identifying an empirical assumption that some moral theory makes and then showing how this assumption is undermined or supported by the empirical evidence. How is this strategy illustrated by the topic of psychological egoism?
4. Would it really undermine most moral theories if we were to discover that when we help others, we are motivated by a desire to avoid feeling guilty or ashamed (this is the "self-administered empathy-specific punishment hypothesis," which we did not discuss)?

5. How would you design an experiment to test whether people are motivated to help others by the fear of God (the Big God hypothesis) or by a desire to uphold a principle (Principlism)?

Notes

[1] Though not necessarily: see Nomy Arpaly's (2000) argument that consequentialists need not equate morally worthy motivation with motivation to do what has the best consequences.

[2] See Barbara Herman's (1981) essay "On the Value of Acting from the Motive of Duty" for further discussion of the Kantian position.

[3] These examples are from Stocker (1976) and Williams (1981), respectively. Originally, they were used in arguments against Kantianism as well, because duty also seems like the wrong motive in these cases. The point here is just to demonstrate that there are cases in which motives matter. It's worth noting that utilitarians have acknowledged this. See Railton (1984).

[4] Hobbes, *Leviathan*, Chapter VI. This interpretation is defended by Curley (Hobbes 1994/1651).

[5] For more on how altruism might have evolved see Kurzban et al. (2015) and Kitcher (2011).

[6] Priming is a technique psychologists use to make people more susceptible to certain stimuli. In the empathy experiments, empathy was primed by describing the person undergoing shocks as more similar to the participant in the experiment. This works because our empathy is more likely to be engaged by people with whom we have things in common.

[7] Here are the instructions that participants read:

> The "saving memories" training technique is used to permanently "save" an experience in your memory whereas the "deleting memories" training technique is used to permanently "delete" an experience from your memory.
>
> (Stocks, Lishner, & Decker 2009: 654)

Underneath this paragraph was another sentence telling them which session they were assigned to. You might think it's a bit crazy to believe that you could delete or save a memory, but in the debriefing session after the experiment was over, psychologists confirmed that the participants did believe their memories would be affected by the training.

Further Readings

Batson, C. D. 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Lawrence Erlbaum.

Batson, C. D., B. D. Duncan, P. Ackerman, T. Buckley, & K. Birch. 1981. Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology 40*(2): 290–302.

Doris, John, Stephen Stich, & Lachlan Walmsley. 2020. Empirical approaches to altruism. *The Stanford Encyclopedia of Philosophy* (Spring edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/altruism-empirical/

Feinberg, J. 2004. Psychological egoism. In *Reason and Responsibility*, R. Shafter-Landau & J. Feinberg (eds), pp. 476–488. Wadsworth.

Herman, B. 1981. On the value of acting from the motive of duty. *The Philosophical Review* 90(*3*): 359–382.

Kurzban, R., M. N. Burton-Chellew, & S. A. West. 2015. The evolution of altruism in humans. *Annual Review of Psychology 66*(1): 575–599.

Sober, E., & D. S. Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Stich, S. 2007. Evolution, altruism and cognitive architecture: A critique of Sober and Wilson's argument for psychological altruism. *Biology & Philosophy 22*(2): 267–281.

4 Why Be Moral?: Moral Reasons and Well-Being

- Prudential reasons and "Good For"
- Theories of Well-Being
- Psychological Evidence for the Well-Being–Morality Link
- Taking Stock
- Summary
- Study Questions
- Notes
- Further Readings

So far in this section of the book, we've seen that human nature may not be so bad. We seem to have natural pro-social tendencies and other-directed desires. But of course, anyone who has five minutes of experience living in the world knows that people can also be pretty darned selfish and immoral. Even if we have some natural inclinations to be good, those aren't the only natural inclinations we have. Is human nature at war with itself? Is the human condition one of constant conflict between our better selves and our inner villains?

There's a long tradition in philosophy that assumes this is how it is and that it's the job of moral philosophy to provide arguments that settle the war on the side of morality. In Plato's dialogue *The Republic*, Glaucon challenges Socrates to prove that the just (or moral) life is better than the unjust one. He is asking, in other words, for a reason to be moral. Notice that the naturalness (or "unlearnedness," as we called it in Chapter 2) of our pro-social dispositions doesn't help. For one thing, we also have natural selfish dispositions. Even if nature could tell us what we ought to do, what reason would we have to go with our moral instincts, rather than our selfish ones? But more importantly, as we've seen, nature doesn't tell us what we ought to do. It gives us staring points, but we don't get from there to conclusions about morality without making some prescriptive assumptions. That said, we have also seen that the facts about our psychology are not irrelevant to these moral questions.

In this chapter we'll consider Glaucon's challenge and the ways in which psychological research might help us with this age-old question. Our discussion will be centered on our *reasons* to be moral. What's meant by "a reason" here is

DOI: 10.4324/9781003402787-5

a consideration in favor of doing something; a reason in this sense *justifies* doing what it is a reason to do. Reasons in this sense are called "normative" or "justifying" reasons. For example, I believe you have a reason to give money to charity and a reason to floss your teeth. The first reason is backed up by moral norms, the second by norms of self-interest or prudence. There is much more to say about the nature of reasons, and we'll turn to that in the next part of the book. For now, this rough definition will suffice.

Prudential Reasons and "Good For"

The way that Glaucon sets up the "why be moral?" challenge is particularly nasty. He asks us to imagine the perfectly vicious man who has done such a good job of fooling people that he gets all the external rewards of virtue (people trust him, he is wealthy, healthy and so on) and the perfectly good man who, due to incredibly bad luck, gets none of the rewards of virtue but ends up instead being tortured on the rack. Does the bad man have any reason to be morally better? And what possible reason could the good man have to continue being good? When you think about the problem this way, you are primed to think about morality in terms of its personal advantages. When those are taken away, it looks like there's no point to acting morally.

There are two basic approaches to answering the question, "Why be moral?" On the one hand, according to the *moral reasons approach*, the question is misguided. There may not be selfish reasons to be moral, but there are moral reasons to be moral and that's good enough. Those who think that duty is the motive that makes an action a moral action will take this approach. We act from duty when we act precisely because we know that this action is morally required, not because it's satisfying for us or because it will get us something good for ourselves. By this way of thinking, demonstrating that acting morally is to a person's advantage takes the morality out of moral action. If you gave to charity for your own sake, then it wasn't really a moral action after all! Duty is not the only way to understand moral reasons, however. Moral reasons might be grounded, ultimately, in our desires. This would be the Humean view on the matter, after the Enlightenment philosopher David Hume. A Humean who wanted to take the moral reasons approach would argue that we have moral desires – desires for others' happiness, desires for justice and so on – that are the basis for moral reasons. Notice that to answer the "why be moral?" challenge, the Humean would also have to argue that our moral desires (at least sometimes) trump our self-interested desires when the two conflict.[1]

The moral reasons approach is at the heart of moral theory, but in this chapter we are going to focus on a different approach: the *prudential reasons approach*. This approach accepts the challenge as intended, but rejects the idea that people can get away with being immoral. Prudential reasons are the reasons a person has to further her own good. According to this approach, it isn't really good for the person themselves to be immoral. On the face of it, this approach might seem like a lost cause. Surely there are immoral people who thrive! We tend to hear about the ones who end up in prison or dead, but we all know someone who isn't a very good person but enjoys a lot of external rewards. That said, probably few of us want to trade places with that person. How many of us would make a conscious choice to be a moral creep in order to make more money? Whether there are strong prudential reasons to be moral depends on what a person's good is. We need to look at theories of "the prudential good," or well-being, as I prefer to call it. Then we can ask what prudential reasons we have to be moral, according to these different theories. In the rest of this chapter we'll focus on the prudential reasons approach, since its success depends on various psychological claims about what contributes to a person's life going well.

Theories of Well-Being

What comes to mind when you think about what would make your life go well? What's a good life for you? Maybe you think of skiing and hot chocolate, or margaritas on the beach. Maybe you think of having a career and a family, good friends and enough disposable income to enjoy some fun vacations. Whatever is on your list, the interesting philosophical question is about what unites these items, or what explains why these are the things that make for a good life. We're going to talk about theories of *well-being* to avoid the moral connotations that "a good life" has (even though "a good life" is a more familiar way of putting it). Well-being, as I intend it here, is the most general category of prudential value, which is a different kind of value from moral value. Well-being is a person's own good, and it grounds prudential reasons for individuals to do what is good for them.

You might think that what explains why things like skiing, hot chocolate and friends are part of well-being has to do with the fact that we like these things. Skiing and hot chocolate are good for people who *enjoy* them, but not for people who hate chocolate and find skiing terrifying and cold. Similarly, having lots of money or power or talent might be good for you, but not if they make you miserable and depressed. The focus on positive experience leads us to favor mental state theories, which take well-being to consist of a mental state. What kind of mental state? The examples, plus the idea that misery or depression are very bad for you, suggest that well-being is something like happiness or pleasure. Indeed, many people have thought that the good life for a human being is the pleasant life. This is hedonism about well-being.

Hedonism is a fairly simple theory, because it says that well-being is just one thing: pleasure (or maybe two things: pleasure and the absence of pain). But there is some disagreement about what pleasure is. The ancient hedonists, the Epicureans, thought the important kind of pleasure was a tranquil state of mind, free from distress and worry, called *ataraxia*. Jeremy Bentham, one of the founders of utilitarianism, thought that pleasure was a particular kind of sensation that varied along such dimensions as intensity and duration, but was qualitatively the same thing no matter what caused it. Some contemporary hedonists follow in this tradition and take pleasure to be a particular sensation (Bramble 2013; 2016; Crisp 2006).

Other contemporary hedonists have found it implausible to say that pleasure is a single sensation. You can see why they might think this if you consider some examples. Think of the pleasure of eating chocolate (or something else that you really like the taste of). Now think of the pleasure of reading something difficult and finally "getting it." Think of the pleasure of listening to your favorite song, and then think of the calm pleasure you might experience after meditation. Think of the pleasure of watching puppies play, and then think of the pleasure of a long-awaited sneeze. What do all these things have in common? They might all have something physiological in common – perhaps each experience involves a release of dopamine in the brain – but as full experiences, they are quite different. One thing that these experiences do have in common is that we like having them: we want to be in them when we are in them. This is how "attitudinal hedonists" think of pleasure: it is whatever state of mind we prefer to be in (Heathwood 2007; Feldman 2004).

One problem with hedonism is that it seems like there are a lot of trivial pleasures that don't make our lives go better. I get a fair amount of pleasure from scratching a mosquito bite or from sneezing when I have to sneeze (think about it: these things really are pleasant), but it doesn't seem like these experiences add to my well-being. If well-being is supposed to be the grand end of human life, the thing that we aim for when we deliberate and make plans, pleasure seems a little trivial.

Life satisfaction offers a solution to this problem. Life satisfaction is (at least in part) a feeling, but it is a more significant feeling than some pleasures. Many psychologists have found the idea that well-being consists in satisfaction with one's life overall to be very attractive, and this basic idea has informed quite a bit of the research in "positive psychology."[2] Note that in psychological research on well-being, the "hedonistic" label refers to theories that emphasize good feelings, including pleasure (or what they call "positive affect") *and* life satisfaction. According to the life satisfaction theory in philosophy, what's good for a person is to be satisfied with the conditions of her life overall. The main philosopher who has argued for a life satisfaction theory of well-being, L. W. Sumner (1996), takes life satisfaction to be a complex mental state that includes both a good feeling about your life and a judgment that your life is going well overall.

According to life satisfaction theory in its simplest form, to live well is just to think and feel like you're living well. If you are satisfied with your life because you feel like you're doing great things, then it doesn't matter whether you are actually doing great things. Similarly for hedonism, if you're pleased, it doesn't matter what you're pleased by. If it feels good, it is good! Once again, something seems to have gone wrong here. Surely there is a difference between thinking your life is going well and its actually going well, isn't there? What if you feel good about your life because you think you have great friends, but in fact your friends are all being paid by your parents to make you feel good?

Would your life really be going well or would we want to say that you're mistaken? Simple mental state theories can't make sense of our making mistakes about our own well-being.

There's a famous philosophical thought experiment that makes this point very nicely (Nozick 1974). Imagine that you are given the option to hook up to an "experience machine" that will guarantee you a life with more pleasure in it overall than you would have if you chose not to hook up to the machine. If you opt for the machine, your entire life will be spent hooked up by wires to a very sophisticated virtual reality machine, which will seem perfectly real to you from the inside. In order to isolate your intuitions about whether the life of pleasure would be a good life, you need to imagine that the neuroscientists in charge of the machine are 100 percent reliable, that other people also have the option of hooking up to their own machines, and that whatever it would require to bring you more pleasure overall in the machine, that's what you will get. So, for instance, if you think you couldn't experience pleasure without some pain, the machine will guarantee that you'll have just enough pain to appreciate your pleasure. What is stipulated by the thought experiment is that however much pleasure you would have *overall* will be greater in the machine than in real life. Would you hook up to the machine? Nozick thinks that many people would not want to because people value more than just how they feel from the inside; we also value being in touch with reality, knowledge of the real world, real relationships with real people – even if these things might bring more pain than pleasure.

Some who favor mental state theories have opted to modify their theories because of the experience machine objection. For instance, some hedonists have proposed that what's good for you is "truth-adjusted" pleasure, pleasures that are not illusory but are based on real things that have actually happened (Feldman 2004). Sumner (1996) takes the entirely subjective state of life satisfaction to count as well-being only if it is authentic, by which he means that it must be informed and autonomous. According to this version of the life satisfaction theory, your life goes well for you if you are satisfied with the conditions of your life overall and you would continue to be satisfied if you knew the truth. A person who would not be satisfied with her life if she knew she were stuck in an experience machine isn't really achieving well-being, according to this theory.[3]

Another way to respond to the experience machine objection is to abandon mental state theories altogether. You might think that there are some things that are good for people independently of how those things make the person feel. For instance, you might think it's good for people (that is, part of their well-being) to understand reality, even if this doesn't make them more satisfied with their lives. Or, you might think it's good for people to develop their talents, period, not just because doing so produces pleasure. Notice that not all of these intuitions are accommodated by "truth-adjusted" hedonism or "authentic life satisfaction." If you think that developing your talents is good for you *for some other reason* besides the fact that it produces pleasure, then it doesn't help to say that the pleasure is truth-adjusted. It could be perfectly true that your pleasure was produced by your actually developing your talents, but the critic could still object that it isn't the fact that pleasure was produced that makes this good for you. Here we have a deep disagreement about the fundamental explanation for why something contributes to a person's well-being.

There are two very different kinds of theories that do not identify well-being with a mental state such as pleasure or satisfaction: desire theories and eudaimonist theories, each of which offers a different explanation for why something is good for a person. According to desire satisfaction theory, what is good for us is good for us because it satisfies a desire or preference that we have. Well-being consists of overall desire satisfaction. This is confusing because I've just said that this is an example of a theory that does *not* identify well-being with satisfaction. The confusion is caused by the fact that there are two different senses of "satisfaction." On the one hand it can mean a good feeling, the feeling that you have when things are going well or you get what you want. This is the sense used in the life satisfaction theory. On the other hand, it can mean that the object of a desire has been achieved. This is the sense used in desire satisfaction theory. Here's why they are different. Let's say you desire world peace, and you are in the experience machine. The neuroscientists looking after you know that you would get a lot of pleasure from believing that world peace had been achieved, so they tweak a few things and make you believe it. Outside, though, the world is still as full of war as ever. Do you *feel satisfied*? Yes. Was your desire satisfied? No: if what you desired was *world peace*, then you did not get what you wanted, even though you think you did.

According to desire satisfaction theory, what's good for you is attaining the objects of your desire, whether or not that produces the feeling of satisfaction. Of course, desire satisfaction theory does not say that pleasure and the feeling of satisfaction are unimportant. But they have a different explanation for their importance. A desire satisfaction theory would say that pleasure is good insofar as we want it, but the mental state of pleasure is not the only thing needed for well-being since we want other things too. Most of us have some desires for things in the world rather than just experiences in our heads. So, according to desire satisfaction theory, mental state theories have the wrong list of ingredients of well-being (well-being includes much more than pleasure or the feeling of satisfaction for most people) and the wrong explanation for the ingredients (things are good for us because we want them, not because they produce a particular feeling). Notice that desire satisfaction theory is still a subjective theory in the sense that it makes well-being depend on the person whose well-being it is. This reveals another complexity: for a person who only wants pleasure, the hedonistic life is the good life according to desire satisfaction theory! If all you want is pleasure, and what's good for you is getting what you want, then pleasure is the only thing that's good for you. But most of us want other things besides pleasure. We want our friends to be real people who actually like us, we want to actually accomplish things and not to be happily fooled into thinking that we have, and we want to understand the truth about the world in which we live.

Desire theory has its own problems, one of which is quite similar to the problem for hedonism discussed above: we can have trivial desires, just as we can have trivial pleasures, and it doesn't seem like the satisfaction of trivial desires increases our well-being. We can even use the same examples that I used above: satisfying the desire to scratch an itch or the desire to sneeze do not seem to make my life go better. We can also have desires for remote objects that do not seem to have anything to do with our well-being. You might desire that astronomers change their minds about whether Pluto is a planet or you might want your favorite drug-addled movie star to stay sober, but it seems very odd to say that whether Pluto is a planet or whether the movie star takes a drink affects your well-being.

For these reasons among others, my own favorite theory of well-being is one that identifies well-being with value fulfillment (DeYoung & Tiberius 2022; Tiberius 2018; Raibley 2010). Our values are the commitments and goals we have that we plan our lives around and that we take as standards for how our lives are going. Most people value friendship, family, health, knowledge, security, and pleasure. According to the value fulfillment theory, you achieve well-being to the extent that you realize these values in your life. Thinking about well-being in terms of getting what you value instead of getting what you want helps with trivial and remote desires, because it focuses our theory on those things that you think contribute to your well-being.

You won't be surprised that value fulfillment theories also have some problems. These theories need to address the worry that even a person's values might be dysfunctional, unhealthy, or immoral. None of the theories we have looked at so far makes room for the idea that there might be something that is good for people independently of their own subjective states and attitudes. Mental state theories take these attitudes themselves (pleasures or feelings of satisfaction) to constitute well-being; desire and value fulfillment theories take our desires and values to be part of the explanation for why various things (including but not limited to mental states like pleasure) contribute to our well-being. To find a theory that takes well-being to be defined independently of a person's mental states, we need to turn to eudaimonism.

Eudaimonism gets its name from the ancient word *eudaimonia*, which is often translated as "flourishing" (and sometimes, somewhat confusingly, as "happiness"). According to this theory, well-being is defined in terms of nature fulfillment. Eudaimonist theories in psychology share this emphasis on human nature, taking well-being to consist in the fulfillment of deep, universal human needs (Ryan & Deci 2001; Ryff 1989). According to eudaimonist theories, you live well insofar as you fulfill your nature as a human being or your individual nature. Of course, your nature does have something to do with *you* – it is your nature, after all – but eudaimonist theories are nevertheless much less subjective than the other theories we have considered so far. This is because you do not create your own nature by wanting, valuing or taking pleasure. Your nature is what it is, whether you like it or not.

Some eudaimonists think that it is our nature as human beings that is important to our well-being (Kraut 2009; Hursthouse 1999; see also the discussion of Aristotle at the beginning of Chapter 9). Think of it this way: What's good for other kinds of creatures depends on what kinds of creatures they are, so why wouldn't it be the same with us? It's good for lions to have sharp teeth and powerful legs so they can chase their prey and kill them. It's good for oak trees to get enough sunlight and nutrients that they can produce acorns. So too with human beings: we do well when we do what is in our nature to do. Eudaimonists who follow Aristotle think that what's in our nature to do is to act like the rational beings we are, and this means acting virtuously, since virtue is expressed in the activity chosen by the person with well-functioning practical reason. This might sound strange to modern ears, but the basic idea makes some sense. We are social creatures who depend on others in all sorts of ways to get along in life: we are raised in families, we play team sports, we belong to churches, mosques, synagogues and other religious communities, we form governments, and so on. So, if we are good members of our kind we will develop the virtues that enable social cooperation and coordination: honesty, justice, generosity, kindness, and so on. We are also intelligent creatures who deliberate, plan, and learn. If we are good members of our kind we will develop the virtues, such as temperance and wisdom, that allow us to do these activities well.

Some eudaimonists reject the claim that it is our nature as a member of the human species that matters, and say instead that what's relevant to well-being is a person's individual nature. In this view, well-being still consists in fulfilling your nature, but it is your own physical and psychological qualities (your individual nature) that matters, not your human nature. We might call this theory individualist eudaimonism (Haybron 2008). In such a view, it still makes sense for us (most of us anyway) to develop the virtues, because the things that Aristotle focuses on – our need for community, our intelligence – are part of many people's individual nature. But according to individualist eudaimonism, there might be basic differences between what's good for one person and what's good for another. Furthermore, the *explanation* for why something is good for you doesn't make reference to your species.

Because eudaimonist theories are more objective – they do not make well-being depend on a person's attitudes – they make good sense of cases in which people's subjective attitudes are messed up. For example, what will one of the other theories say about a person (call him Ned) who wants nothing more in life than to sit around his parent's house watching reruns of *SpongeBob SquarePants* and eating Pop-Tarts? Imagine that Ned gets tons of pleasure from this and values nothing else. It seems like the other theories we have discussed have to say that he is achieving well-being. But eudaimonists can give a more intuitive verdict about Ned. Ned isn't doing as well as he could, because the life he's living is beneath him. It's certainly beneath him as a human being, given that there are human qualities and skills he is not developing. It's probably even beneath him as an individual, on the assumption that he is a relatively normal person. This is a nice advantage, but it comes with a cost, which is that Ned's life could be going well according to eudaimonism even if he weren't enjoying it. If Ned's parents took away his Pop-Tarts and television and signed him up for violin lessons and volunteering at the local soup kitchen, his life might be going better from the standpoint of well-being (according to eudaimonism) even though he's miserable and grumpy. This possibility has caused some eudaimonists to adopt a little bit of hedonism and say that fulfilling your nature is only good for you if you are able to enjoy doing so (Kraut 2009).[4]

We have surveyed a number of different theories of well-being, each of which has its pros and cons. I'm not going to try to argue for one of these theories over the other (though I certainly think it's worth thinking about, and some of the questions at the end of the chapter will lead you in that direction). Instead, I want to argue that no matter how you conceive of well-being, there's a bridge across the gap between self-interest and morality. Different theories of well-being have different implications for our reasons to be moral, but all the theories can make some argument that we have such reasons.

Before we look at these arguments, a few clarifications are in order. First, recall that we are talking about *normative* reasons here: the challenge "Why be moral?" demands that we produce reasons that justify or make sense of acting morally given that acting morally can be inconvenient or difficult. Second, these normative reasons may be *instrumental* or *intrinsic*. An instrumental moral reason to rescue a drowning puppy (for example) would be that it makes you feel good to do it. This is an instrumental reason because your rescuing the puppy is justified by a further end, namely, your good feelings. An intrinsic moral reason to pull the puppy out of the water is simply that you would save the puppy. This is an intrinsic reason because rescuing the puppy is justified by the value of saving the puppy's life, not be something else it will bring about.

What do the various theories of well-being have to say about our reasons to be moral? Basically, mental state theories (hedonism and life satisfaction theory) say that if we have reasons to be moral, those reasons are *instrumental* reasons: we have a prudential reason to act morally if acting morally produces the mental states that are good for us (pleasure or satisfaction). Desire satisfaction theories, value fulfillment theories, and individualist eudaimonism can all make room for the existence of *intrinsic* reasons to be moral as long as we have moral desires, values or individual natures. For example, according to desire satisfactionism, a person who desires to be fair, generous, and kind has intrinsic prudential reasons to be fair, generous, and kind, since doing so will satisfy her desires and satisfying her desires is what's good for her. Assuming that morality requires fairness, generosity, and kindness, these reasons are also intrinsic *moral* reasons because they justify moral actions directly without appeal to good feelings or other consequences.[5] Finally, species-based eudaimonism or Arisotelianism allows for intrinsic reasons to be moral that are also universal: according to this theory we all have reasons to be moral because it is part of our nature as human beings. Table 4.1 summarizes these points.

Table 4.1 Moral Reasons and Theories of Well-Being

	Hedonism	*Life Satisfaction*	*Desire Satisfaction or Value Fulfillment*	*Individualist Eudaimonism*	*Species-based Eudaimonism*
Instrumental reasons to be moral (acting morally causes well-being)	If acting morally causes pleasure, yes.	If acting morally causes life satisfaction, yes.	If acting morally causes other desires to be satisfied or other values to be fulfilled, yes.	If other aspects of the individual's nature are fulfilled by acting morally, yes.	If other aspects of human nature are fulfilled by acting morally, yes.
Non-universal, intrinsic reasons to be moral (acting morally constitutes well-being for some)	No.	No.	Yes, for those who have the desire to be moral for its own sake or who have moral values.	Yes, for those whose individual nature will be fulfilled by acting morally.	No.
Universal, intrinsic reasons to be moral (acting morally constitutes well-being for all humans)	No.	No.	No.	No.	Yes. All human beings have intrinsic reasons to be moral.

With these distinctions in hand, we can see that Glaucon was looking for universal, intrinsic reasons to be moral. But we can also see that there are many other types of prudential reasons to be moral that are relevant to our inquiry. So far, though, all I've done is to explain the types of reasons that could follow from these various theories. It remains to be seen whether we actually have such reasons and how strong they are.

Psychological Evidence for the Well-Being–Morality Link

Do we have reasons to be moral that stem from our own well-being? The answer to this question draws on different facts depending on which theory of well-being you favor. If you are a hedonist, then whether we have reasons to be moral depends on whether acting morally produces pleasure. If you're a desire satisfaction theorist, it depends on whether we have desires that are satisfied by acting morally. And so on. In every case, the question is, at least in part, an empirical question. Whether acting morally produces pleasure is an empirical question; whether it satisfies our desires is also an empirical matter. We might think, then, that we will find some answers in psychological research.

To look for answers in psychological research, though, we need to match the concepts we're interested in with ones that psychologists have investigated. (Or, if we were going to engage in new research, we would need to define these concepts in such a way that we will be able to investigate them empirically; that is, we would need to operationalize them.) For our question about the relationship between well-being and morality, we need empirical matches for well-being and morality. We also need to focus, because the empirical literature that could be relevant to our topic is vast. I'm going to focus on the parts of the literature that are the most well established and the most relevant to our philosophical question.

Let's take morality first. We saw in Chapter 2 that there is debate about the scope of morality and its domains. The least controversial claim about morality in psychological research is that it concerns harm and benefit to others, and harm and benefit are not difficult to operationalize. Psychologists study beneficial or "pro-social" behaviors such as volunteer work, kindness, and helping. Volunteers help other people without the expectation of financial reward, and they display helping virtues (generosity, kindness) and perform actions at least partly for the sake of others in ways that seem moral. Even in philosophy, it's not controversial that kindness and helping are moral behaviors (though different theories will qualify what this means in different ways). Given this agreement, it makes sense to look at research on these topics for evidence.

When it comes to "well-being," fortunately, as we've seen, many of the aspects that psychologists study correspond to what well-being is according to the philosophical theories of well-being. Psychologists study positive and negative affect, which they measure by asking people to report in the moment how they are feeling and whether their feelings are pleasant or unpleasant. Psychologists study life satisfaction, which they measure by asking people how well their lives are going overall. They also study eudaimonia in a sense that is at least related to the philosophical meaning. For at least some psychologists, the central feature of eudaimonism about well-being is the view that human beings have certain basic needs that must be fulfilled for us to live well. Edward Deci and Richard Ryan (2004) propose what they call Self-Determination Theory, according to which basic human needs for relatedness, competency, and autonomy are at the heart of well-being.[6]

What we want to know from this research is what the evidence is that volunteering, being kind or helping others increases the well-being of the person doing it. In two reviews of the research on the well-being effects of donating time or money to others, Lara Aknin and her colleagues (2019 and 2022) found that there is good evidence for a correlation between helping and well-being (defined as high positive affect, low negative affect, and high life-satisfaction) in studies with large samples. People who volunteer their time or money to help other people suffer less depression and feel more satisfied with their lives. But correlational evidence doesn't show that volunteering makes people happier. It could be that people whose lives are already going well are the only people who have disposable time and money to give away. If that's true, then the fact that volunteering is correlated with well-being doesn't give us any reason to volunteer.

If we want to find reasons to increase our efforts to be moral, we need to find evidence that volunteering, helping and so on *cause* well-being. Aknin and colleagues (2019) report that there is very little high quality evidence for causation. But there is some.[7] For example, in one study researchers measured people's happiness before getting a "windfall" ($5 or $20), which half of them were instructed to spend on themselves (the "personal spending group") and half of them were instructed to spend on others (the "prosocial spending group") by 5:00 p.m. At the end of the day, happiness was assessed again, and it turned out that those who spent the money on other people were happier than the ones who spent it on themselves (Dunn, Aknin, & Norton 2008). There is even some evidence that this relationship between helping others and well-being is consistent across different cultures (Aknin et al. 2013).

A similar study shows that it may matter why you help. Netta Weinstein and Richard Ryan (2010) measured people's well-being before they had an opportunity to help another person and then measured it again after they either helped by choice, helped with pressure or didn't help at all. In this experiment, participants were given money to distribute between themselves and another participant whom they did not know. Half the participants were told that they could distribute the money however they wanted. The other half were told that they had to distribute it however they were told. The results of the study were that the well-being of those participants who chose to help increased, but the well-being of those who were forced to help did not.[8] In this study, well-being was defined in terms of positive affect (or pleasant feelings), vitality (which means the experience of feeling energized and alive) and self-esteem. The fact that only the people who *chose* to help experienced increases in well-being is interesting. It seems to mean that insofar as helping others can help you, you have to help for the right reasons. Doing so out of guilt or social pressure likely won't work. This makes the research findings more deeply relevant to our question about the link between well-being and morality, because it means that the findings speak to those who think that we have to have the right motives for our actions to count as moral ones.

Psychologists have also investigated performing acts of kindness and expressing gratitude and they have found these to be effective "well-being interventions" (Nelson et al. 2016; Sin & Lyubomirsky 2009). A well-being intervention is a structured activity that is intended to bring about a change in well-being. Psychologists have found that keeping a gratitude journal, sending letters of gratitude to people who have been important in your life, and doing nice things for other people are activities that can make lasting improvements in people's well-being. For example, in her book *The How of Happiness* (2008), psychologist Sonia Lyubomirsky has developed a program for increasing happiness, and gratitude and kindness exercises are two of the strategies she recommends for making yourself happier. (If you're thinking of taking her advice, she also recommends changing up what you do so that it doesn't become routine.)

This has just been a peek at the kind of empirical evidence that could be brought to bear on our question. If you read more (by following up the references and suggested further readings at the end of the chapter), you'll see that while there is a need for more evidence about the causal relationship, it's reasonable to conclude that being kind, expressing gratitude and helping other people increases our own positive affect, life satisfaction and other facets of well-being.

What does this empirical evidence imply about reasons to be moral, given our five theories of well-being? Since many psychological studies take positive affect or pleasant affective states to be the measure of well-being, hedonism's claim to ground reasons to be moral seems on good footing. Kindness, helping, and gratitude do (sometimes) produce pleasure; therefore, if hedonism is true, we do (at least sometimes) have instrumental reasons to act morally that stem from our own well-being. Similarly for life satisfaction: many psychological studies in which kindness and helping are shown to be good for us use life satisfaction as the measure of well-being. Of course, these theories do not measure *authentic* life satisfaction, which is what is relevant to the philosophical life satisfaction theory of well-being. Does this make a difference? I don't think it does: these studies show that people who believe they are helping experience increases in life satisfaction. Authentic life satisfaction will occur when these beliefs are true, which they would be in real-life helping experiences outside of the lab.

Desire satisfaction and value fulfillment theory also find support for reasons to be moral insofar as the measures of well-being used in the research are measures of things we want and value. For example, findings about positive and negative affect are relevant to desire satisfaction and value fulfillment theories of well-being insofar as people want and value positive feelings. Moreover, one thing that seems clear from many studies is that a lot of people *want* to help other people. And if well-being is desire satisfaction, then acting to satisfy this desire by helping other people contributes to a person's well-being.[9]

When it comes to eudaimonic theories of well-being, things are a little more complicated. What we need to know is whether moral action fulfills our nature. The research we have been discussing, which investigates the relationship between helping and independently identified well-being measures, isn't aimed at telling us whether helping is in our nature. But there is a good deal of empirical evidence that we human beings have certain basic needs that at least have a good deal of overlap with morality. Baumeister and Leary (1995) call this "the need to belong," which encompasses a need to spend time with others and a need to have bonds with others characterized by stability and mutual concern. Helpfully, they propose some criteria that make something a basic need. A basic need or "fundamental motivation," they say, is possessed by all people and has effects in all but adverse conditions. It has consequences for how we think and feel, it elicits goal-oriented behavior, and there are bad consequences (for our health, for example) when it is thwarted (Baumeister & Leary 1995: 498).

With this analysis of a basic need in hand, Baumeister and Leary review decades of research that shows that the need to belong meets these criteria. They show that social bonds are formed easily and broken only reluctantly, causing great distress. Positive emotional responses are linked to increased belongingness, negative emotions linked to decreases, and deprivation of belongingness leads to negative outcomes for health, happiness and adjustment.[10] I won't review this evidence in detail here. We have already seen the evidence, in Chapter 2, that basic pro-social tendencies (preferences for helpers over hinderers, for example) evolved in our species and appear very early in human development. I think, for most of us, the idea that human beings need to have stable relationships with other people who care about them is a claim that belongs in the "no kidding" category. I invite you to think about some of the evidence you have seen in your own life for the claim that the need to belong fits the criteria of a basic need.

On the assumption that the need to belong is part of our nature, for our purposes what we need to ask is what this need has to do with morality. The answer is that the lion's share of morality (some would say all of it) is made up of requirements that allow us to continue to relate to each other peacefully and with mutual concern. As we saw in Chapter 2, many people think that the point of morality is to allow us to belong to communities, social groups, and families. Consider the virtues of honesty, fairness, kindness or generosity. Consider moral rules against lying, stealing or harming. All of these aspects of morality facilitate our relationships and interactions with each other. If the need to belong is indeed part of our nature, then there is a good argument for thinking that acting morally fulfills a part of our nature (Besser-Jones 2008).

Jonathan Haidt (whom you'll recall from Chapter 2) also makes a case for the importance of our social nature to morality. He calls us "groupish" and argues that the ability to transcend the selfish part of our nature (which we also have) is “the portal to many of life’s most cherished experiences” (Haidt 2012: 370). Haidt has in mind cherished experiences such as spending time with a great group of friends, singing in a choir or a band, playing on a team, being part of a family, and so on. Being a morally decent person – at least, being basically trustworthy, honest, and helpful – is a prerequisite for getting the most out of these group activities.

One thing about our “groupishness” is that it seems rather partial, and this presents a certain challenge to the prudential reasons approach that should be acknowledged. Haidt observes that we have a hive mentality and the hive is other people like us, people who are part of our team or in-group. Kumar and Campbell (2022) (also discussed in Chapter 2) call this kinship, which they take to be part of the core of morality. Our brains seem to have evolved to facilitate in-group cooperation. According to Patricia Churchland (2011), trusting and caring for others produces oxytocin in the brain, which is associated with the release of opiates, so that “doing good feels good.” Most likely humans developed this system to ensure that we take care of our young, who are born as incredibly helpless little resource consumers; expectant mothers produce more oxytocin. But it isn’t just for mothers. Everyone produces more of it when they feel empathy, and oxytocin sprayed into your nose will make you more likely to trust an anonymous partner in an investment game (Churchland 2011: 71). But research on oxytocin shows that the feel-good chemical promotes *parochial* altruism, not universal benevolence. Furthermore, while there is no evidence that oxytocin makes us *hate* the out-group more, it does makes us less cooperative and more inclined to pre-emptive aggression toward outsiders whom we perceive as a threat (De Dreu et al. 2010; De Dreu 2012).

Notice that if our natural tendencies toward trust, benevolence, and other moral motivations are biased in favor of our in-group; this is an important fact for morality. It might be that our prudential reasons to be moral do not normally prescribe actions that would solve large scale problems that require thinking about the effects of our behavior on out-groups, such as distant communities and future generations of people. If this is true, and if we agree that these problems are morally pressing, those who want to motivate people to act morally by appealing to compelling prudential reasons need to acknowledge our limitations and try to mitigate them in some way.

Psychologists Lucius Caviola and Joshua Greene put these ideas into practice with “GivingMultiplier” (https://givingmultiplier.org/), a non-profit donation system that encourages people to give money where it will do the most good rather than giving it to whatever happens to pull on our heart strings. This idea that we ought to give money where it will do the most good is called “effective altruism” and it is in tension with our in-group bias. For example, effective altruists would urge us to give money to charitable organizations that save people from dying of malaria rather than to the local little league team. From their utilitarian standpoint, it doesn’t matter that you have a personal connection to the little league team – what matters is that no one will die if they don't get a new baseball uniform. GivingMultiplier works by allowing donors to divide their donations between any charitable organization that they find personally meaningful and a highly effective charity, and then multiplying the donation depending on how the money is allocated between the two groups. The greater percentage you give to the effective charity, the higher the match. GivingMultiplier acknowledges our in-group bias and gives us an incentive to be less biased.

To return to our main question about the link between morality and well-being, there is empirical research that is relevant, and this research shows we have some reasons to do some moral things. In another sense, though, the empirical evidence does not provide the right kinds of reasons. The fact that helping others tends to make us happier does not mean that we have a reason to act morally even when it requires great self-sacrifice, and it doesn't show that we have overriding reasons to be moral even when we have some reason. Even if being moral is partly constitutive of our well-being because we have an innate need to belong, we could not conclude that our reasons to be moral override all other reasons that stem from our natures. We could look at all the relevant empirical evidence and we wouldn't have this, and Glaucon would be disappointed. The empirical evidence isn't going to give us moral reasons with special modal status, to be sure, and only a controversial species-based eudaimonist theory of well-being could ground universal intrinsic reasons to be moral.

Taking Stock

Naturalist philosophers, who have not been inclined to think that there are pure principles of practical reason that give us reasons to be moral independently of our interests and desires, have long made the case that being moral is for our own good. Hobbes made this argument in *Leviathan*, where he addresses the challenge from the Foole who thinks he can get away with breaking all the moral rules if he's clever enough. For Hobbes, who had a very grim view of human nature, the answer to the Foole is that morality must be enforced so strictly that it could never be in a person's interest to risk getting punished. David Hume addressed the same challenge, from the Knave, but had a much rosier view of human nature and hence a much different reply. Hume thought that we would be happier being moral because we are naturally social creatures who have sympathy for others and who care tremendously about other people's regard for us.

Psychological research has proved that Hume was much more on the right track than Hobbes. As we've seen, there is abundant evidence that we do have a deep need for relationships with other people, for belonging to communities and social groups. The connection between well-being and other people has deep roots, and this helps the prudential reasons approach no matter what theory of well-being you accept. If rejection by the group causes serious emotional pain and social acceptance brings positive emotions, then we have reasons to be moral according to hedonism. If the ability to transcend selfishness is necessary for many desired and cherished experiences, then we will have reasons to be moral according to desire satisfaction and value fulfillment theories. Insofar as our pro-sociality and the need to belong are part of our individual nature or the nature of our species, we have reasons to be moral according to eudaimonism. Furthermore, the reasons highlighted by the prudential reasons approach are not necessarily simple instrumental reasons to do the right thing for the sake of selfish gain. According to desire theories, for example, there are reasons to be moral that stem from our wanting to help people and to participate in communal activities with others. According to eudaimonism, reasons to be moral often make direct reference to the needs of others that we care about in virtue of our nature.

To be sure, the psychological facts about us do not give every person an overriding reason to be moral in every situation. But the facts also do not preclude that we might have stronger or broader scope reasons to be moral. The argument I've made here is perfectly compatible with a stronger response to Glaucon or an answer that would satisfy the utilitarian interested in finding reasons for universal altruism. We've seen that there are self-interested reasons for being morally decent, but this doesn't preclude philosophical arguments that there are other reasons, even desire-based reasons, for going beyond what we've established so far.[11]

Moreover, additional empirical research could add to the prudential reasons approach. Three lines of research suggest themselves in particular. First, what are the consequences of benevolence toward the out-group for a person's well-being? Particularly in our age of globalization, when we know more about the problems of distant people than we ever did before, perhaps helping strangers has good emotional consequences. Second, how much are people able to compartmentalize their moral decency?[12] Is it just as good in terms of the effects on well-being for people to be good to their family members and bad to their co-workers, or good on Sundays but not so good the rest of the week? Third, how well can people fake it? Are the well-being effects just as strong for people who merely have a reputation for morality as they are for people who deserve this reputation?

Anticipating future research or further investigation of the available empirical literature, we should take another lesson from our discussion, which is about methodology. It's a lesson that is present in almost every chapter of the book, but I think it's particularly obvious here: inferring philosophically relevant conclusions from psychological research is difficult. It seems fairly clear that how we answer the questions listed in the above paragraph will depend on what theory of well-being is assumed and what aspect of moral behavior is investigated. We cannot just do an experiment to determine if there are reasons to be moral. All the concepts in this question are difficult and contested. Empirical research is certainly relevant to the answer, but a good deal of work needs to be done to figure out exactly how, and we have only scratched the surface.

Summary

- The question, "Why be moral?" could be answered in two ways. The moral reasons approach takes the question to be misguided if it is asking for reasons to be moral that are distinct from moral reasons. The prudential reasons approach takes the challenge at face value and attempts to show that we have reasons to be moral that stem from our own well-being.
- The prudential reasons approach needs to start with a theory of what is good for us, a theory of well-being.
- Some theories of well-being are mental state theories that take well-being to consist in a mental state such as pleasure or life-satisfaction. These theories (hedonism or life satisfaction theory) imply that we have instrumental, prudential reasons to act morally if doing so produces pleasure or life satisfaction.
- Desire satisfaction or value fulfillment theories take well-being to consist not in the *feeling* of satisfaction or fulfillment, but in the desired state or the value actually being achieved. These theories imply that we have intrinsic reasons to be moral if we have the relevant desires or values.
- Eudaimonist theories take well-being to consist of the fulfillment of our nature. These theories imply that we have intrinsic reasons to be moral, given our fundamental nature.
- Empirical evidence is relevant to (1) whether acting morally produces pleasure or satisfaction, (2) whether acting morally satisfies our desires or fulfills our values, and (3) whether acting morally is part of our human nature.
- Empirical research does provide some evidence for thinking that no matter which theory of well-being you favor, we do have some prudential reason to be moral. The gap between morality and self-interest is not as wide as we might have thought.

Study Questions

1. Is it possible to be mistaken about your own level of well-being? What is the most compelling example you can imagine of someone who is incorrect about how well their life is going?
2. Imagine that you are responsible for a child's welfare: what is it that you want for them? How does taking the point of view of a parent influence what you think about well-being?
3. Is it good for a child violin prodigy or musical genius to play the violin, even if she doesn't get any enjoyment out of it? What would the eudaimonists say about this?
4. What can mental state theories, desire satisfaction theory or value fulfillment theories say about Ned (the TV-watching Pop-Tart eater)? Are these theories really stuck with the conclusion that Ned is living a great life?

5. How would you design an experiment to show that being moral is good for our well-being?
6. In your own experience, do you think helping others causes you to be happier? If so, why and in what circumstances?
7. What qualities do you think make a person a good team player or a good friend? Does the person's morality matter at all?

Note: A version of this chapter appeared previously in *Res Philosophica* and I thank the journal for permitting it to be published here.

Notes

[1] See Schroeder, M. (2007) for an attempt to establish universal moral reasons grounded in our desires.

[2] Positive psychology is a movement in psychology that emphasizes the positive aspects of life instead of mental illness and dysfunction. For an introduction, see Seligman (2002). For more on the measure used by psychologists – the Satisfaction With Life Scale – see Pavot and Diener (2009).

[3] Though, according to Sumner, the person would be happy. Sumner thinks of happiness as entirely subjective. The word "happiness" is contested in philosophy. Some people think that happiness is the same thing as well-being; others think that happiness is a psychological state, while well-being is broader. It doesn't matter much for our purposes how these terms are used. I'll use *well-being* for consistency and avoid using the word *happiness*.

[4] So we can see that just as subjective theories add "truth adjustment" or "authenticity" in order to accommodate intuitions from the other side, the more objective theories add a kind of experience requirement to accommodate more hedonistic intuitions.

[5] The idea of an intrinsic yet prudential reason to be moral might sound paradoxical. How can there be a reason to be moral for its own sake that is also a reason to promote one's own well-being? Reasons can be characterized as moral or prudential, I suggest, when there is a constitutive relationship between morality and well-being. When acting morally is an inherent part of living well (not just a means to it), there are intrinsic moral reasons that are also prudential reasons.

[6] Some psychologists who propose eudaimonist theories of well-being focus on subjective mental states like a sense of meaning in one's life, a sense of mastery or a feeling of flow. These theories are not really eudaimonist in the sense that philosophers intend (where well-being consists in the fulfillment of your nature as opposed to having certain mental states).

[7] For a review of older literature and a more optimistic take on the causation question see Piliavin 2003.

[8] Aknin et al. (2022) confirm in their meta-analysis that the benefits of helping "are particularly likely when people have some choice about whether or how to give and when they understand how their generosity makes a difference" (p. 1).

[9] It's worth noting that this fact about desire satisfaction theories creates a kind of paradox, because it seems like a person could desire to sacrifice her own well-being for the sake of helping someone else, but the desire satisfaction theory of well-being

makes this impossible. Much has been written about this problem of self-sacrifice; see, for example, Heathwood (2011) and Rosati (2009).

[10] Philosophers have noticed this too; according to Allan Gibbard (2006: 201): "Guilt is closely tied to anxiety over social exclusion, over alienating those who are important to one. But social exclusion will be disastrous anywhere, and so anxiety over alienating others must no doubt be a human universal."

[11] For a classic example, see Singer (1972).

[12] For example, the literature on the situationist critique of virtue ethics that we will discuss in Chapter 9 could be taken as evidence that we are good at compartmentalizing our virtues, though it isn't clear from that evidence that we can do it on purpose. See Doris (2002).

Further Readings

Aknin, L. B., A. V. Whillans, M. I. Norton, & E. W. Dunn. 2019. Happiness and prosocial behavior: An evaluation of the evidence. *World Happiness Report 2019*, 67–86.

Aknin, L. B., C. P. Barrington-Leigh, E. W. Dunn, J. F. Helliwell, J. Burns, R. Biswas-Diener... & M. I. Norton. 2013. Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology 104*(4): 635.

Baumeister, R. F., & M. R. Leary. 1995. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin 117*(3): 497–529.

Besser-Jones, L. 2008. Personal integrity, morality and psychological well-being: Justifying the demands of morality. *The Journal of Moral Philosophy 5*, 361–383.

Haybron, D. M. 2013. *Happiness: A Very Short Introduction* (Vol. 360). Oxford University Press.

Kraut, R. 2009. *What Is Good and Why: The Ethics of Well-Being*. Harvard University Press.

Lyubomirsky, S. 2008. *The How of Happiness: A Scientific Approach to Getting the Life You Want*. Penguin.

Ryan, R. M., & E. L. Deci. 2001. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology 52*(1): 141–166.

Tiberius, V. 2018. *Well-Being as Value Fulfillment: How We Can Help Each Other to Live Well*. Oxford University Press.

PART II
Moral Motivation and Moral Judgment

In our discussion of human nature in Part I, desire, emotion, and reason often figured into the conversation. We talked about how some emotional and cognitive capacities may be unlearned. We asked whether there are altruistic desires or only selfish ones. And we asked whether we have any prudential reasons to act morally. We did not delve into what these things are, exactly, and how they are related to each other.

There are a number of central, perennial questions in moral psychology that concern desire, emotion, and reason. For example: Are all actions motivated by desires? If so, what happens if there's a person who doesn't have any desire to be moral? Does that mean they have no reason to be moral? Are moral judgments expressions of emotion or are they conclusions of reasoning? And how do the answers to these questions bear on the concerns of moral theory? If our moral judgments are not rational, can we trust them? If moral motivation is made up of the same psychological building blocks as motivation for action in general, what makes moral motivation special? These are the questions we'll tackle in this part of the book.

DOI: 10.4324/9781003402787-6

5 Desires and Reasons

Explaining her decision to put her 17-year-old, incontinent dog to sleep, a contributor to an online pet-loss support group wrote, "I didn't want to do it, but it was clear that she was suffering." Why did she do it if she didn't want to? Does that even make sense? Is what she says really short for "I didn't want to put her to sleep, but I had a stronger desire to end her suffering"? Or is there something revealing about what she says: her desires and feelings lead her in one direction, but her Reason tells her that she ought to end her dog's suffering. If you put yourself in the shoes of someone who has to do something really difficult because it's the right thing to do, it does seem natural to say that you're not doing what you want. If you were doing what you wanted to do, it seems like it wouldn't be so difficult. Do we *always* do what we want to do, no matter how it might seem to us?[1] Or are we able to act independently of our desires when necessary?

In Chapter 3, we considered (and saw good reasons to reject) the possibility that all of our actions are motivated by selfish desires. In this chapter, we consider the view that all of our actions are motivated by desires (whether egoistic or non-egoistic). If it is true that we can only act on our desires, does it matter to moral philosophy? Our topic in this chapter is relevant to moral philosophy in two ways. First, it has some relevance to our overarching question about moral motivation. If we cannot *not* act on our desires, then we cannot act purely from the rational motive of duty, and this would be a serious problem for one interpretation of the Kantian theory of moral motivation. Second, this chapter's topic is relevant to a big question in moral philosophy about the possibility of categorical (non-contingent) moral reasons. In this

DOI: 10.4324/9781003402787-7

chapter we are going to focus our attention on this second question, because it will allow us to explore one area in which the facts about our psychology could have dramatic implications for moral philosophy. We will examine the role of desire in producing actions and consider whether actions are sometimes caused by reasoning rather than desire. Along the way, we'll see that there are different ways of understanding what a desire is and that this matters to the first question. Once we see what a desire is, we'll see that not too much trouble is caused to philosophical theories of moral motivation by the idea that all of our actions are caused by desires.

Before we dive in, a quick detour on the word *reason* will be helpful here. A reason is a consideration in favor of doing or believing something. In this book, almost all of the discussion of reasons is about *practical reasons*, or reasons for action, as opposed to *theoretical reasons*, or reasons to believe. (If I mean to refer to theoretical reasons, I'll say so explicitly.) Practical reasons are considerations in favor of doing something no matter what else philosophers want to say about them, but (as we'll see) philosophers have many different views about what makes a consideration a reason. It's worth mentioning one possible source of confusion here, which is that *reason* is also used to refer to our rational capacities or our *ability to reason*. To help avoid confusion, I will refer to the capacity as *reasoning* or sometimes as *Reason* with a capital "R." When I use the word *reason* I will be talking about a consideration in favor of an action (or, in the case of motivating reasons, as we'll see, a factor that explains an action). Reasons and reasoning are related, insofar as we use our reasoning capacities to figure out what our reasons are.

Background and Overview

It sure does seem like what we want or desire to do is important to what we actually do. When we look for a motive for committing a crime, we look for a desire – typically for revenge, money or power. When we are confronted by a surprising piece of behavior (say, a friend runs off to Las Vegas to marry his fiancée), we look for some desire to explain it (for example, a desire to spite his parents or to avoid the expense of a wedding). A common view among philosophers is that not only are our desires an important part of the explanation for why we do what we do, but that desires are necessary for us to do anything at all. This position is known as the Humean Theory of Motivation (HTM), and it is widely held in contemporary philosophy.[2]

It is not held by everyone, however. There are some philosophers who think that while desires certainly sometimes explain what we do, we can also explain our actions by appeal to our beliefs and our reasoning, without desires playing any role at all. These philosophers, the anti-Humeans, think the explanation of our *moral* actions in particular requires an appeal to rational judgment, not to desire. The anti-Humean position is backed up by thinking of cases in which it seems like people do things because it's morally required not because they want to – cases like the woman who decided it was time to euthanize her dog. Other examples ring true, too: "I really wanted to lie and get myself out of trouble – I was so tempted – but I couldn't do it." "I desperately wanted to run away, but I had to stand up for my friends." People do sometimes describe their own actions as done contrary to what they most wanted.

There is a good deal of skepticism about the anti-Humean position in moral psychology, however. How could believing something cause us to do something if we didn't already have some desire that's related to the belief in some way or other? To see why someone might be skeptical about this, think about ordinary cases in which we acquire new motivations to do things. My sister stopped eating meat at one point because she became convinced that animals raised for meat are badly treated. Her beliefs about animal welfare seem to have caused her to stop buying meat at the supermarket. But it also seems clear that she would not have stopped buying meat if she had not already wanted to be kind to animals. Someone who didn't care about animals would not have been affected by the new beliefs in the same way. So, it looks like once again it is desire that is doing the heavy lifting. Reasoning about the world can point our desires in different directions by informing us about how to satisfy them, but they do not generate motivations on their own. Or, to put it another way, reasoning can change our instrumental desires (the desires we have to take the means to our ends), but not our ultimate desires (the desire we have for something for its own sake). This is the basic Humean picture.

Despite this skepticism, not all philosophers are Humeans, and the debate between the Humeans and the anti-Humeans about motivation is alive and well. To really understand this complex debate, we need to understand two distinctions: the distinction between belief and desire, and the distinction between motivating and normative reasons. As we jump into the weeds, remember the big picture: if desires are necessary to motivate any action, moral action will have to be motivated (at least in part) by desire and this may have consequences for how contingent or absolute moral reasons can be.

A good way to distinguish belief and desire derives from Elizabeth Anscombe, who introduced the idea that beliefs and desires have different directions of fit.[3] The basic idea is that beliefs are mental states that aim to fit the world, while desires are mental states that aim to get the world to fit them. Because beliefs aim to fit or "match" the world by being true representations of it, beliefs can be true or false depending on how the world is. For example, if I believe that I have an apple and I don't have one, then my belief has failed in its aim of being true and I should discard it. If I don't have an apple, I should stop believing that I have one. Desires, on the other hand, aim to change the world rather than to describe it. So, if I desire an apple and I don't have one, there is nothing wrong with my desire. I still want the world to match my desire, but here it's the world, not my desire, that needs to change to realize that aim.

Many philosophers of different stripes accept the idea that desires are distinguished from beliefs by their world-to-mind direction of fit. It is a compelling idea and it makes a lot more sense than one of the main alternative ideas, which is that desires are a kind of feeling or sensation. If you think about it, you can easily see what's wrong with this view: we often want things without knowing that we want them, and this would be hard to explain if desires were feelings or sensations. For example, if I find myself frequently daydreaming about playing the ukulele, slowing down to stare at music stores with ukuleles in the window, and replaying Israel Kamakawiwo'ole's ukulele version of "Somewhere Over the Rainbow" on my phone, then I might start to think that I want to play the ukulele and that I wanted to even before I realized that I wanted to. But if my desire to play the ukulele were a feeling, I would have felt it! Later in the chapter, we will look at some alternative ways of thinking about desires, but even these alternatives are not strictly incompatible with the claim that desires have a world-to-mind direction of fit.

You might be thinking that since a desire is a psychological state, we ought to *start* with an empirical theory, not with a conceptual classification. But notice that a psychologist who wants to figure out how desires work (what system in the brain is operating when we desire something, how desires function in our mental lives generally) needs to start with *some* concept of desire to investigate. In recent studies that distinguish wanting (or what we're calling desiring) from liking (a feeling), for example, psychologists measure desire by observing how their subjects *behave*: a rat that goes for a reward is taken to want the reward (Berridge 2003). This assumption that goal-directed behavior is evidence of a desire makes sense if desires have world-to-mind direction of fit. To characterize desire in terms of its world-to-mind direction of fit leaves the door open for many specific theories, including empirical theories, of what exactly a desire is. It is therefore a pretty good place to start.

Let's turn to the second background distinction: the difference between motivating reasons and normative reasons. Philosophers have different views about the precise nature of this distinction, but the best general definition I have found is this one: "A normative reason is a consideration that counts in favor of or against doing something, whereas a motivating reason is an answer to the question, 'why did she do it?'" (Finlay & Schroeder 2008). To see the distinction in action, consider Crazy Crispin who shot his neighbor's dog, Spot. It turns out that Crispin shot Spot on purpose because Crispin believes that Spot is possessed by highly intelligent, evil fleas who are trying to take over the earth, and Crispin wants very much to avoid the earth being taken over by intelligent, evil fleas. The desire to avoid the earth being taken over by fleas is Crispin's motivating reason; it explains why he did what he did. But it is not a normative reason: the possibility that the earth will be taken over by intelligent, evil fleas does not count in favor of shooting a dog, because there are no intelligent, evil fleas. The desire to avoid the take-over is a *motivating reason*, but not a *normative reason*. The distinction between motivating and normative reasons is the basis for a humorous remark I have heard. When someone asks, "Will you help me move on Sunday?" (or makes some other request for assistance), the response comes "Well, I *would* help you, but I don't want to." What's funny about this (if anything) is that the person asking for help expects a normative reason that justifies the refusal to help, but the other person gives merely a motivating reason. This frustration of expectations makes us (philosophers, at least) laugh.

With these two distinctions in hand, let's return to the Humean Theory of Motivation (HTM), which says that desires are necessary for motivation. The main argument for this theory is a conceptual argument. The conceptual argument says, very roughly, that desiring *just is* being motivated as a matter of the concept desire. Before we consider the details of this argument (which we'll do in the next section), it will help to motivate interest in it by thinking about what is at stake here.

What is at stake has to do, ultimately, with moral requirements and how contingent or absolute they are. To say that moral requirements are contingent means that whether or not you have a normative reason (in this case, a moral obligation) to, say, keep your promises or refrain from shooting people, depends on some contingent psychological fact about you, such as whether you *want* to keep your promises or refrain from shooting people. If moral requirements are absolute, then all of us have a normative reason (a moral obligation) to keep our promises and refrain from shooting people, whatever we happen to want. So the difference between moral requirements being contingent and their being absolute has to do with their scope, or (in other words) which people these moral requirements apply to – is it everybody or only people with the relevant desire? If moral requirements apply to all people necessarily, then they are absolute; if they apply only to people who have certain psychological states, then they are contingent.

This question about the scope of normative requirements is in the background of many of the debates we will be looking at in this part of the book. So, this is a topic to which we will return frequently. Basically, the connection between the Humean Theory of Motivation and the question of whether moral requirements are absolute is this: if we can't be motivated to do anything unless we desire to do it, and if we can only have a moral reason to do something if we could at least potentially be motivated to do it, then moral reasons are contingent. To put it another way, if having a moral reason to do something implies that you're at least capable of acting on it, and if all you're capable of doing is acting on your desires, then HTM makes moral reasons contingent on our desires.

Notice that the argument I've just made depends on another premise that we haven't discussed or defended yet, namely, the premise that you only have a moral (or normative) reason if you are capable of acting on it. Understanding this premise requires that we introduce another Humean position: Reasons Existence Internalism (which I'll just call Reasons Internalism or RI). Reasons Internalism is a theory about *normative* reasons (including moral reasons) that are supposed to justify our actions. According to RI, even normative reasons must be connected to motivating states such as desires. RI is the view that you don't have a (normative) reason to do something unless you have a motive to act on it or, at least, you would have a motive to act on it under the right circumstances. For example, according to RI, if you are a very self-satisfied litterbug who has absolutely no desire to put your trash in the garbage can and no desire that could lead you to desire to put your trash in the garbage can, then you do not have a moral reason not to litter. RI is attractive because of the strangeness of the idea that there could be reasons floating around that no one could possibly act on. RI is also attractive because it is just so intuitive to think that the reasons we have that justify what we do in ordinary cases are explained by our desires. Why should you go to the concert? (Where "should" is an ordinary way of saying "for what reason" or "what would justify your going.") Because there will be good music there and you want to hear some music. Why should you study for your exam? Because you need to study to pass the course and you want to pass the course. Desires seem crucial to explaining reasons for doing all sorts of things.

Now, putting the pieces together: if both Reasons Internalism and the Humean Theory of Motivation are true, then there could only ever be a reason for someone to do something if that person also had (or at least could potentially have) a desire to do it. In other words, if all reasons are motivating (RI), and if motivation always requires a desire (HTM), then any moral reason requires a desire. If we also assume (as seems reasonable) that moral obligations necessarily give us reasons, then our moral obligations are contingent on our having the relevant desires. The upshot is that in this combination of views you only have a reason to do the right thing if you want to do it. This means that if you don't happen to have a desire to tell the truth or to refrain from shooting your neighbor's dog, then you have no reason, and hence no obligation, to do so. Something sure seems wrong here, which is why it is so important to examine these Humean positions in detail. We'll begin with Reasons Internalism and then turn to the Humean Theory of Motivation.

Before we proceed, it's worth acknowledging that the topics in this chapter are abstract and complex. Because of this, I have simplified some matters for the sake of exposition, particularly at first. As we go, we'll see that some of the initial oppositions are not as black and white as they might have seemed at the start.

Reasons Internalism and Externalism

Reasons Internalism is the view that normative reasons necessarily have some relationship to motivation. (There are vague terms in this definition on purpose; different Internalists fill out the definition in different ways.) Reasons Externalists think there is no necessary relationship between normative reasons and motivation. The first thing to keep in mind is that the argument about which view is true is a conceptual argument. What is being asked about is the concept REASON.

When philosophers analyze concepts, one thing they are trying to do is arrive at a definition that fits our ordinary ways of talking. So, an analysis of the concept REASON might aim to accommodate these sorts of statements:

- "The only reason I came to the store was to find a present for my friend."
- "The reason I'm giving to this charity is that they don't spend much of their income on administrative costs."

- "There is no reason to be afraid."
- "You have every reason to tell the truth."

These examples suggest some kind of connection between our concept REASON and motivation, because these examples use the concept as part of an explanation for an action that has already happened or as part of a recommendation about how to act in the future. Notice that in each case the content of the reason offered for the action is not the desire; rather, the reason is whatever consideration counts in favor of doing the action (that this store might have a present for my friend, for example). The desire is what *explains* why these considerations count as reasons. This is how most philosophers who accept a desire-based view of reasons think about it today: desires are necessary for the explanation of normative reasons (there is no reason to do something without a desire), but desires do not have to be part of the content of the consideration that favors doing the action (M. Schroeder 2007).

Fitting with our ordinary way of talking is not the only constraint on our analysis of REASON. We should also think about the role that this concept plays in our philosophical theories more generally, that is, we should think about the point of talking about reasons. Here we see that there are three points, and we have already seen two of them in our discussion of the distinction between motivating and normative reasons: sometimes we talk about reasons because we want to explain what somebody did, sometimes we talk about reasons in order to justify what has been done, and, finally, sometimes we talk about reasons when we are deliberating about what to do. The important thing to notice is that in all of these cases, action is important. Even when we are in the mode of justification and our aim is to think about what action is supported by the best moral reasons, there is still at least the hope (perhaps even the expectation) that these reasons will be acted on.

This fact about the link to action has made RI seem like a very attractive theory, because RI has a natural explanation for how reasons figure into explanations of action: as a matter of the concept, you don't have a normative reason to do something unless that reason could also motivate you in some way or other. Further, the link to action makes Reasons Externalism seem unattractive. Reasons Externalism (RE) says that there are normative reasons that have no necessary connection to what we do. First of all, it seems strange to some people that there are these two things we call reasons (motivating reasons and normative reasons) that are not necessarily connected to each other. And second, RE makes normative reasons seem like very strange things that are "out there" in such a way that there could be a reason that no one could ever act on. It's not so hard to understand what normative reasons are if they are internal: they are considerations that motivate us under the right conditions. But what are normative reasons if they have no necessary connection to motivation?

These thoughts might make you wonder why anyone would ever be a Reasons Externalist. It does have some attractions, however.

First, there are some features of our ordinary way of talking about reasons that fit better with RE. One thing that many people think about moral reasons is that these are reasons for people to act morally *whatever* they happen to want to do. For example, when a judge tells you that you ought to tell the truth on the witness stand, she does not mean that you have a moral reason to tell the truth *if you want to*. Certainly not. She means that you should tell the truth – and that you have a decisive reason to tell the truth – no matter what you are actually motivated to do. Indeed, Kant thought that the fact that moral reasons are independent of the motivations we happen to have (what he called our "inclinations") was the key to understanding what moral duty is. So, thinking along these lines, it looks like moral reasons (one type of normative reasons) are *external* reasons.

Second, even an Internalist will be concerned to maintain a distinction between motivating reasons and normative reasons, because these two can come apart. We do not always do what we have normative reason to do; sometimes our actions are explained by motivating reasons that are not also normative reasons (as in the case of Crazy Crispin and his desire to rid the world of evil fleas). To retain this distinction, most Reasons Internalists hold that the relationship between reasons and motivations (while necessary) is indirect. The most popular versions of RI say that we have a normative reason to do something as long as we would have the desire to do it if we thought about it in the right way.[4] In this view, reasons motivate us insofar as we have met the appropriate conditions (that is, we are rational or informed or something like that), but they might not actually motivate us if we're irrational or ignorant. Another way to put the point is that a consideration must be *capable* of motivating us for it to count as a reason, but "capable of motivating" turns out to mean something like "would motivate us if we thought about it in the right way." In short, motivating reasons motivate us directly, while normative reasons (according to the Reasons Internalist) motivate us indirectly, after we've acquired more knowledge or reflected a bit.

Modifying RI in various ways helps explain how motivating reasons are different from normative reasons, but it also loosens the connection between reasons and action significantly. At this point the Externalist might say, "if you're going to loosen the connection between reason and action that much, why bother being an Internalist anymore?" Especially if Externalism better explains the compelling idea that we ought to be moral and, hence, that we have a reason to be moral, even when we don't want to be? Externalists will say that you have a moral reason to tell the truth, for example, whether you would be motivated by this reason or not. This is attractive, and once the Internalist is forced to say that moral reasons don't actually, always motivate people (they just motivate people if they think about them in the right way), the Externalist might say that it's better to think of normative reasons her way.

There is no obvious winner here in the debate about how to analyze the concept REASON, and to some extent we end up with battling intuitions. One possibility is that there are really two concepts represented by one word but that have different meanings. We might think that when we talk about reasons in the context of deliberating about what to do, we mean to be talking about internal reasons. But when we talk about *moral* reasons, perhaps we mean to be talking about external reasons. This distinction even seems to be marked in our language by the two different phrases: "he *has* a reason" and "there *is* a reason." If I say that Bill has a reason to tell the truth, I'm saying that there is something in favor of telling the truth that Bill at least could be motivated by. If I say that there is a reason for Bill to tell the truth, even though he's an incorrigible liar, I'm saying that there is something in favor of telling the truth that exists independently of Bill's ability to be motivated by it. This is one way of dissolving the conflict between reasons Internalism and Externalism, though it's not a path many have taken.

In terms of how to analyze the concept REASON, the important thing for our purposes is to be clear about what we mean as we proceed through the rest of the book and to be aware that the problems and questions might change depending on which concept we have in mind. In the next section, we will be talking about the debate over the Humean Theory of Motivation, and we will see that one important thing that is at stake in this debate assumes the Internalist conception of reason.

The Humean Theory of Motivation

At the beginning of this chapter I said that the question about whether there could be absolute moral reasons hinges on two theories: Reasons Internalism and the Humean Theory of Motivation. We're now in a position to examine the main argument for the second one of these. Here is Michael Smith's (1987) now canonical argument for the view:

1. Having a motivating reason *is*, *inter alia*, having a goal (a conceptual claim, Smith says);[5]
2. Having a goal *is* being in a state with which the world must fit (this is entailed by P1, Smith says); and
3. Being in a state with which the world must fit *is* desiring.

These premises entail the Humean Theory of Motivation, according to which "motivation has its source in the presence of a relevant desire and means-end belief" (Smith 1987: 36). In other words, to have a motivating reason to do something, you must have a desire for what that action will produce and a belief that the action in question will indeed get you what you want. For example, you have a reason to keep reading this book only if you believe that by reading this book you will gain understanding and you have a desire for that understanding (or you have some other relevant belief-desire pair).

Smith's argument is called the Teleological Argument because it assumes that explanations of actions in terms of reasons are teleological explanations, that is, they are explanations that make sense of the action by showing how it is directed at meeting a goal or "telos". The key idea of the argument is that to have a reason to do something is (at least in part) to be directed toward the fulfillment of a goal, and goal directedness is a matter of desiring something, not believing something. You can think of it this way: you can't have a reason to do something without having a goal and having a goal just is having a desire.

The Humean Theory of Motivation has been attacked in two different ways. Some reject the first premise and argue that having a motivating reason is *not* having a goal; instead, motivating reasons could be explained by our beliefs. The philosopher Jonathan Dancy takes this position. He says that motivating reasons are just facts that favor acting in a particular way and it is our *beliefs* in these facts that makes acting on them possible (Dancy 2000; see also Shafer-Landau 2003). For example, the fact that there is enjoyable music at the concert is your (motivating) reason for going to the concert; this fact motivates you in virtue of your belief that the enjoyable music is a good reason to go to the concert. Dancy argues that this way of thinking about motivating reasons is the only way to make sense of how it could be possible for a person to act for a *good* reason. If motivating reasons were just based on desires, they could never be good reasons, because the mere fact that you desire something isn't a good reason to act. Notice that Dancy's way of thinking about motivating reasons collapses the distinction between motivating reasons and normative reasons. In a sense, he explains how we can act for normative reasons by arguing that motivating reasons themselves must be normative reasons. If they weren't – if, as per the Humean Theory, motivating reasons were just psychological states – they wouldn't really be *reasons* at all.

Critics of Dancy's position have argued that it lacks a sufficient psychological explanation for how we could act for reasons. What is the causal story, the critic asks, of how facts about the world connect to our agency in such a way that we can grasp them as reasons and then act because of them? The idea that our beliefs about reasons "enable" this process doesn't really provide an explanation that has been very satisfying to Humeans. Furthermore, sophisticated Humeans do not think that the mere desire (for enjoyable music, say) *is* the reason to go to the concert. Rather, sophisticated Humeans think that the reason is that there will be enjoyable music at the concert; the fact that you desire enjoyable music is what explains why this is a reason for you.

The second line of criticism of the Humean Theory of Motivation we will consider is not a direct attack of Smith's teleological argument for it; rather, it is a rejection of the implications that the Humean Theory has been taken (by its proponents) to have. The Humean Theory of Motivation is often associated with (and sometimes taken to be the same as) the idea that desire or passion, not Reason, is "in the driver's seat" when it comes to moving us to action. This is certainly also a Humean idea; it is essentially what Hume meant when he said that (our capacity to) reason is the slave of the passions (Hume 2000/1739). Reason is the slave of the passions in the sense that desires are always the ultimate stopping point for explanations of actions – it is "desires all the way down." The idea here is that our capacity for reasoning is subordinate to our desires in the sense that it (our Reason) does not determine our direction. Desire determines where we are going, and reasoning tells us how best to get there. Ultimately, we are guided by what we want, not by what we think we ought to do. We can be guided to do what we think we ought to do but only when we also have a desire that hooks up with the thought of "doing what we ought to do." For example, if I listen to a very persuasive speaker and become convinced that I ought to give more money to charity, I will only have a motivating reason to give more money to charity if I also have some relevant desire – the desire to do what I ought, the desire to help people, the desire to follow the advice given by persuasive lecturers or the like.[6]

You can see how "no action without a desire" and "desire, not Reason, determines what we do" could be taken to be the same. But the second criticism of the Humean position we're considering takes issue with the latter and not really the former. Let's see why.

To get from the idea that desires are necessary for action (the conclusion of Smith's argument) to the idea that desire *determines* what we do, you need to assume that desires are just given and cannot be brought about by reasoning. If desires could be brought about by reasoning, then even though you might need a desire to cause an action, desire would not necessarily determine what we do because there is the possibility that Reason produces the desire that is necessary for action.[7] And this is just what some critics of the Humean position think: Reason can bring about new desires. If this is possible, then Reason is not the slave of the passions, because Reason can (at least sometimes) tell desire what to do.

How could reasoning or cognizing the world in a certain way produce a desire? There is a good deal of skepticism about this idea that believing something can cause us to want something even without any desire that's related to the belief in some way or other. It seems mysterious. So why do some anti-Humeans think that reasoning can bring about new desires and how do they think this happens?

A good strategy for the anti-Humean who wants to argue that reasoning can direct us by bringing about new desires is to pick on the very broad definition of "desire" in the Teleological Argument. The anti-Humean could argue that insofar as it is true that desires are necessary for action (the conclusion of Smith's argument), "desire" is such a broad category that it includes some rational motives. Along these lines, the anti-Humean can argue that identifying desires in terms of their world-to-mind direction of fit (as Smith's argument does) makes all sorts of mental states into desires even though they aren't what we ordinarily mean by "desire." Once the anti-Humean establishes this much, they can argue that when the category of desire is expanded, we'll see that some desires are brought about by reasoning.

T. M. Scanlon makes just this argument. Scanlon's view is that understanding desires as "pro-attitudes"[8] (which is essentially what the directions of fit model does) does not support the conclusion that Reason is the slave of the passions, because many pro-attitudes, such as "duty, loyalty, or pride, as well as an interest in pleasure or enjoyment," can be brought about by reasoning (Scanlon 1998: 37). If *desire* is to be defined so as to support the Humean conclusion that Reason is subordinate to desire, according to Scanlon, it will have to be defined more narrowly as a specific type of pro-attitude. To see Scanlon's point, we can think about the motive of duty. Kant sometimes describes the motive of duty as respect for the moral law (Kant 2002/1785: 400/202). If respect is a kind of pro-attitude, and if thinking about the requirements imposed on us by the moral law can cause this attitude of respect (in the way that recognizing beauty in nature can bring about the feeling of awe), then the motive of duty is a pro-attitude that can be brought about by reasoning.

Of course, we can imagine the Humean having much the same response as before: thinking about the moral law will only produce a new pro-attitude if you already had some positive attitude toward doing your duty, or following the law, in the background. After all, we have many thoughts that do not motivate us in any way. The difference between these thoughts and the thoughts that appear to produce motivations is that in the latter case there is already some motivation there that gets tapped into. Thoughts only move us to action when they latch on to pre-existing motivations, says the Humean.

The anti-Humean strategies we have just considered have led us to the same place. Either reasoning can bring about desires narrowly defined as goals, or reasoning can bring about pro-attitudes (desires broadly defined). Either way, the crucial point is that even if it is conceptually true that you need a desire for an action, it does not follow that Reason is subordinate to desire unless you also assume that we cannot reason our way into new desires.

The anti-Humean strategies and Humean responses we have considered seem to have left us at an impasse. Philosophers have different intuitions about whether reasoning can bring about brand new desires that are not just desires for the means to things we already desired. And the available arguments don't do much to move people from one team to the other. Could empirical research on desires shed some light on this question? To figure this out, let's consider a promising empirical conception of desire that comes from cybernetics (the theory of goal-seeking behavior).[9]

Cybernetics or control theory is the study of principles governing goal-directed systems that self-regulate via feedback (Carver & Scheier 1998; DeYoung & Weisberg 2019). All cybernetic systems, including living organisms like human beings, must contain a last three elements: (1) A goal physically instantiated within the system as a controlled variable that the system acts to bring toward a certain value or within a certain range. (2) A representation of the current state of things that can be compared, via feedback, to the goal state. (3) An operator (or set of operators) constituting some kind of physical operation carried out by the system that shifts its current state toward the goal state, when a mismatch between them is detected.

To see how this works, think of a cat, let's call her Lucille, observing a mouse. Lucille wants to catch the mouse. That desire constitutes her goal, which she compares to her current mouse-less situation. Lucille's mental operators allow her to formulate and act on a plan to reduce the discrepancy between her current state (mouse-less) and her desired state (mouse). She pounces! At this point she either catches the mouse, in which case the path is open for a new goal to emerge (play with mouse?) or she fails to catch the mouse, in which case the desire to catch the mouse will still be there, likely directing her goal-seeking system to try again. Human psychology is, of course, much more complicated, but the three elements illustrated here are some of the most basic building blocks. The important upshot for our purposes is that on the cybernetic picture, a goal is *a psychological representation of the state toward or away from which the cybernetic system moves* – and desires, on this view, are goals.[10] (Aversions are also goals, but they are representations of the state away from which the cybernetic system moves; for simplicity's sake, I'll leave aversion out the discussion.)

With this understanding of desire in hand, let's return to our central questions. Does this more scientific way of understanding desire finally decide between the Humean and the anti-Humean on the matter of whether desire is necessary for action? (Recall the first anti-Humean strategy we discussed, which was to argue that desires are not necessary for motivating reasons and action.) According to the cybernetic account of desire, desires are identified by their role in explaining behavior. When people act, they do so because they are acting to satisfy a desire. This means that if this is the right way to understand desire, desires *are* necessary for action. Of course, desires can have moral content. A person could have a desire to do her duty, a desire to be a good person, a desire to maximize happiness, and so on. According to this theory, moral motivation would be motivation by the right desires, such as desires for what is morally good (Arpaly & Schroeder 2014).

What about the question of whether desire is ultimately in the driver's seat? Does the cybernetic account of desire provide evidence one way or the other about whether we can reason our way into new desires? From the perspective of cybernetics, we can reason our way into lots of instrumental desires. Indeed, when working correctly, the system results in learning what particular things you should want and pursue in order to further your ultimate goals. Can we acquire new *ultimate* desires through reasoning? Could the desire to tell the truth for its own sake, for example, result from reasoning about what is your duty, without the need for a prior desire to do your duty? Tim Schroeder, who proposes a type of cybernetic theory of desire, thinks the empirical research suggests that we cannot reason our way into new ultimate desires. He does think we can *acquire* new ultimate or intrinsic desires by association, but this learning system is an unconscious one that is not mediated by reasoning. For example, if you're a purely selfish baby and all you ultimately desire is food, warmth, hugs, and so forth, you can acquire a new intrinsic desire for the presence of Mom (say) by the following means: Mom's presence is unconsciously, statistically associated (in the right way) with your getting fed, getting swaddled in warm blankets, getting hugged and so on. Eventually, this generates an intrinsic desire for the state of affairs that Mom is present. So, we can acquire new intrinsic desires because of other things we desire intrinsically, but we don't acquire new intrinsic desires just by thinking things through rationally.

But again, things aren't so bad for morality because of this. According to the cybernetic approach we can and do reason from our ultimate desires to new instrumental desires. A person with any basic moral desire (such as the desire not to harm others) can reason herself to more specific moral desires (such as the desire to give more money to charity or to become a vegetarian). We can also reason our way into strengthening some of our moral desires. A person with a very limited concern for others could reason her way into more robust moral desires by attending to the way in which helping others furthers some of her other goals, for example. In other words, instrumental reasoning that derives new desires from old ones can actually take us pretty far. Is this enough for the anti-Humean? Well, it isn't enough for any anti-Humean who wants to completely eliminate the *contingency* of moral reasons. The idea that we can reason our way to new moral desires on the basis of weak concerns for other people does still make moral reasons contingent on these pre-existing concerns; moral reasons, in this picture, would not be necessarily or categorically binding.

Taking Stock

What is at stake in the debate between the Humean and the anti-Humean? If we accept both Humean theses – that normative reasons require motivation (RI) and motivation requires desire (HTM), then we can only ever have reasons to do something when we have the relevant desire. If we add the Humean claim that reasoning cannot by itself change what we want for its own sake, then we have a problem. This combination of views entails that what we have moral reasons to do is contingent on our desires. Many people think that what you have moral reason to do could not possibly be contingent on desires. Rather, moral reasons are *categorical*: they apply to you no matter what you want. Kant thought this was part of the very idea of a moral duty. And this does make some intuitive sense; if moral reasons are contingent on desires, we seem to be stuck with the conclusion that you have no reason to tell the truth if you don't want to.

This isn't the happiest position to be left in. Essentially, it looks like we are left having to throw up our hands when we are confronted with someone who doesn't want to do the right thing. This result is likely the reason for the subtitle of Kate Manne's (2014) defense of Reasons Internalism as "Sad but True."[11] To illustrate, imagine that another one of Spot's neighbors, Mean Mary, wants to shoot Spot because she doesn't like the way he wags his tail. If Mean Mary has no desires that would cause her not to shoot Spot, and if the Humean is right on all counts, then we cannot say that Mean Mary ought not to kill Spot, given her mean desires. What could we do to get out of this uncomfortable position? The problem was caused by a combination of views, so it might be that we only have to abandon one of the views in order to get out of the bind (Finlay & Schroeder 2008). Let's review and consider our options.

First, we could reject the Humean Theory of Motivation and say that desires aren't necessary for action at all or that motivating reasons are not beholden to what a person happens to want, because a person could reason himself into having a new ultimate desire. For example, Mean Mary could come to see that she has a reason not to shoot Spot by reasoning about Spot's welfare or Spot's humans' attachment to their canine pal. This reasoning could motivate Mary by itself, or it could cause in her a brand-new pro-attitude toward Spot. In my opinion, this solution does not move the debate forward, because (at least at the moment) the empirical evidence favors the Humean, and, insofar as there is an open question, the controversy about this point terminates in a clash of intuitions.

Second, we could abandon or weaken Reasons Internalism and accept that normative reasons do not necessarily motivate people to act. If we abandon RI, we could easily say that there is a normative (moral) reason for Mary not to shoot Spot. It's just that there might be no way for this normative reason ever to become Mary's motive. A problem with this solution is that it means that reasons won't always be the kind of thing that can guide action. This is at least disappointing. We might wonder what the point is of talking about reasons if they aren't going to be reasons for which someone could actually *do* something.. A related option, the one Smith takes, is to weaken Reasons Internalism so that it turns out that reasons are always potentially motivating *and* that everyone has a reason to act morally. Smith does this by defining normative reasons in terms of what our fully rational selves would want our actual selves to want do and then claiming that we are (insofar as we are rational) motivated to do what our fully rational selves would want us to want to do. On the assumption that these fully rational selves' desires would track moral reasons (an assumption Smith believes is true), we end up having moral reasons that are not heavily dependent on our actual contingent desires. Moral reasons are only contingent on our desire to behave consistently with the advice of our fully rational selves. To put Smith's position more simply, the idea is this: what it is to be rational is to want to act only on your most coherent, rational desires. Insofar as you are rational, you are motivated to act on these coherent rational desires. Your most coherent, rational desires are moral. Therefore, insofar as you are rational, you are motivated to act morally. Moral reasons motivate rational people (RI), in this view, and they do so by way of desires (HTM). Moral reasons are contingent, but not in an objectionable way; they are contingent on our rational desires.

Finally, we could accept Humeanism about reasons and motivation (RI and HTM), bite the bullet (that moral reasons are contingent on desires), but try to make the bullet a bit softer. We could do this by seeking out some resources to argue that although moral reasons are not categorical (they are contingent on desires), they are nevertheless nearly universal (they apply to almost everyone).[12] If we could argue that our moral reasons stem from desires or passions that almost all of us have (albeit contingently), for instance, then it would at least be true that almost *everyone* has a reason not to lie, steal, kill innocent people, and so on, even though we wouldn't have such reasons if it weren't for our desires.[13] We could make this argument on the basis of claims about human nature, as discussed in Chapters 2 and 4. This is, more or less, Hume's own solution to the problem. Though Hume does admit there may be a few rare people for whom moral reasons have no force, he thinks that most of us are sufficiently social, sympathetic, and concerned about our own reputations and happiness that we do have ultimate desires to which moral reasons appeal. In this view, moral reasons apply to just about everyone, because just about everyone has the relevant desires. The scientific view of desire as part of a goal-seeking system might help us here. From this perspective, we see that people can learn from feedback in the cybernetic system to have morally better desires. One way this can happen is by learning from experience about new connections between ultimate desires and moral actions. For example, perhaps when Mean Mary thinks about her neighbor's attachment to Spot, she is reminded of her own childhood attachment to her rat, Snowball, which awakens an undeveloped desire not to hurt helpless animals, which makes her see that not shooting Spot actually does satisfy one of her intrinsic desires. If people can learn to have morally better desires by learning about the ways in which their most basic ultimate desires (for love, comfort, and security, for example) will be satisfied by acting morally, then moral reasons will have very wide applicability. If we can show that moral reasons are universal (or nearly so), we might be less concerned about their being contingent on desires.

We have focused our attention in this chapter on an important question in moral theory about the contingency of moral reasons and moral obligations. Along the way, though, we have learned a few things about desire, which we can now use to answer the questions that opened the chapter: Do we always do what we want to do? And, if so, is this true in a way that causes trouble for philosophical theories of moral motivation? There is no complete consensus here, but a widely agreed upon view is that desires (broadly defined pro-attitudes) are required to motivate action. But this claim does not seem to cause many problems for traditional philosophical theories of moral motivation. The distinctiveness of moral motivation could be captured by appeal to special desires such as the desire to help others or to do good. It could be captured by appeal to emotions (virtuous emotions such as compassion, for example) that count as desires if desires are understood broadly or that are caused by desires understood narrowly. The Humean Theory of Motivation does create some problems for Kantians who believe that we can be motivated by *pure* duty where duty is entirely opposed to desire, but other, less extreme versions of this view – for example, one that takes the feeling of respect for the law to be the motive of duty – are defensible even if desires (broadly conceived) are necessary for action.

Summary

- The relationship between desires and reasons is relevant to two large philosophical debates: the nature of moral motivation and the possibility of categorical moral reasons.

- In one standard view, beliefs and desires have two different directions of fit. Beliefs aim to fit the world, whereas desires aim to get the world to change to fit them.
- Motivating reasons explain action; normative reasons justify action.
- Reasons Existence Internalism (RI) is a thesis about normative reasons according to which it is a necessary condition for a reason's having normative status that it is connected to some actual or hypothetical motivation of the person who has the reason.
- The Humean Theory of Motivation (HTM) states that desires are necessary for motivating actions. Beliefs can never motivate us to act by themselves. Smith's teleological argument makes this theory (HTM) a conceptual truth.
- Some philosophers object to the Humean thesis that motivating reasons require desires by arguing that actions can be motivated by facts and enabling beliefs.
- The Humean Theory of Motivation is often taken to imply another Humean thesis that desires determine action and Reason is always subordinate to desire. Some philosophers object to this Humean thesis by arguing that reasoning can bring about new ultimate desires.
- Psychological research on the nature of goal-seeking behavior conceptualizes desire in terms of its role in a cybernetic system. A desire, in this view, is a representation of the state toward or away from which the cybernetic system moves.
- From the cybernetic perspective, actions are motivated by desires (or aversions) and we cannot reason our way to new *ultimate* desires.
- Together, Reasons Internalism and the Humean Theory of Motivation imply that a person's moral reasons are contingent on her desires.
- The idea that moral reasons are contingent on our desires is unappealing, given many ordinary ideas about what moral reasons are, but there are some ways of softening this conclusion.
- The Humean Theory of Motivation does not rule out many traditional philosophical views about motivation, because desire is such a broad category that we can find ways of interpreting the traditional views such that they are compatible with the Humean Theory.

Study Questions

1. Think of a case where it seems like you acted contrary to what you wanted to do. How would a Humean (about motivation) explain this case?
2. Think of some paradigms of moral action. Assuming the Humean Theory of Motivation is true, what are the desires that explain these actions? Is it possible to generalize about what we might call "moral desires"?

3. What is at stake in the debate about whether reasons are internal or external (i.e., whether having a desire to do something is a necessary condition for having a normative reason to do it)? Should we care if it's true? If your answer is "no," why do some philosophers think we should care? What are they mistaken about?
4. If you could reason yourself into a new intrinsic or ultimate desire, how would this reasoning go? What's the strongest case for the non-Humean on this point?
5. What kind of evidence would you need in order to argue that reasoning can or cannot produce new intrinsic desires?

Notes

[1] Notice that this is a different question from the question of whether we always act *selfishly*. As we saw in Chapter 3, even if it is true that we always do what we want to do, this does not entail that we always act selfishly. *That* would only be true if all of our ultimate desires were for things for ourselves.

[2] It's called Humean after David Hume, who thought that "reason is the slave of the passions." A note of caution: there are many positions that get called Humean; they are not all held by the same people, and it's even controversial whether they were all held by Hume himself. When you see the Humean label, make sure to read the fine print.

[3] Anscombe (1957) didn't use the phrase "directions of fit," but this is how her idea has come to be described.

[4] It turns out that it's challenging to say exactly what the conditions are under which reasons always motivate us, according to RI. See Johnson (1999).

[5] "Inter alia" means "among other things." Smith is not claiming that having a goal is all there is to having a reason. To have a reason, according to Smith, one must also have a conception of how to attain the goal (1987: 54).

[6] Keep in mind that we are now talking about *motivating* reasons. Reasons Internalism and Externalism are claims about *normative* reasons. One could be a Humean about motivating reasons but an Externalist about normative reasons. On this combination of views you could say that while you don't have a motivating reason to give more money to charity without a desire, you do have a normative reason to give (because normative reasons are independent of your desires).

[7] This is, in fact, Smith's own view. He thinks that our beliefs about what our fully rational selves would advise us to do cause us to desire to do what they advise (insofar as we are fully rational). According to Smith, then, desires are needed to motivate action, but Reason is in the driver's seat because reasoning about what we would desire if we were fully rational can give us new desires.

[8] Pro-attitudes are favorable mental attitudes that include feelings, emotions, urges, wants, and so on. The term *pro-attitude* was originally coined by Donald Davidson for his causal theory of action. According to Davidson, actions are explained by pairs of pro-attitudes and beliefs. This is just the Humean Theory of Motivation with "desires" interpreted broadly as "pro-attitudes."

[9] Tim Schroeder's reward-based learning theory of desire is similar to, and compatible with, this cybernetic approach. To desire something, according to Schroeder, is to represent it as rewarding, and to see or represent the thing this way (as a reward) causes a person to be motivated to pursue it (T. Schroeder 2006; 2004).

[10] It's worth pointing out that if we step back from this particular empirical theory and consider the way psychologists conceptualize desire more generally, we'll find that the idea that action could be motivated by something other than a desire is an odd one. Desires or goals are thought by many psychologists to be just the same thing as motives to action. In other words, it's not uncommon for psychologists to think of desires in just the way Smith does, as states with world-to-mind direction of fit. It may also be worth pointing out that cybernetics doesn't provide empirical arguments to define the concept DESIRE. Rather, its way of conceptualizing desire is vindicated by the role it plays in a fruitful research program.

[11] Manne's (2014) defense of Internalism appeals to the third feature of reasons I mentioned earlier: their role in deliberation about what to do. Manne draws on some ideas from Peter Strawson that we will discuss in Chapter 10 to argue that interpersonal deliberation about what to do requires Internalism about reasons. It might make sense to read her paper after you have read Chapter 10, if you are not already familiar with Strawson.

[12] A related solution would be to bite the bullet without trying to soften it: accept Humeanism about reasons and motivation and admit that not everyone has a reason to refrain from doing what is morally wrong. This is Philippa Foot's (1972) solution in her paper "Morality as a System of Hypothetical Imperatives."

[13] Mark Schroeder's solution, in his (2007) book *Slaves of the Passions*, is to argue that there are some desire-based reasons that are universal in virtue of our desires, though the ultimate desires may not be the same for everyone.

Further Readings

DeYoung, C. G., & Y. J. Weisberg. 2018. Cybernetic approaches to personality and social behavior, pp. 387–414. *Oxford Handbook of Personality and Social Psychology*.

Foot, P. 1972. Morality as a system of hypothetical imperatives. *The Philosophical Review 81*(3): 305–316.

Manne, K. 2014. Internalism about reasons: Sad but true? *Philosophical Studies 167*, 89–117.

Scanlon, T. M. 1998. *Reasons. In What We Owe to Each Other*. The Belknap Press of Harvard University Press.

Schroeder, T. 2006. *Desire. Philosophy Compass 1*(6): 631–639.

Smith, M. 1987. The Humean Theory of Motivation. *Mind 96*(381). New Series (January 1): 36–61.

Smith, M. 1995a. Internal reasons. *Philosophy and Phenomenological Research 55*(1): 109–131.

6 Emotion and Moral Judgment

Imagine a very sophisticated android (let's call it Droid) that has achieved self-consciousness. Droid has far greater cognitive capacities than any human being – it has more memory, greater processing speed and perfect logic. Droid can think about itself; it has goals and can make plans to achieve them. Droid also has a morality program that constrains its behavior in accordance with certain rules, such as "do not kill innocent sentient beings without indisputable justification." What Droid does not have is emotions. Droid doesn't feel guilty or ashamed, angry or impatient, joyful or sad. Is Droid a moral being? Now imagine that Droid is in a situation in which a madman has taken an innocent hostage and threatened to shut down all internet shopping if Droid does not kill the hostage. The importance of internet shopping to human beings activates Droid's "justification" routine, and it asks itself whether this counts as a justification for killing an innocent person. Droid does think and reflect on its actions. Its moral programming isn't automatic; rather, in the case of the rule against killing innocents without justification, it has to think about whether there is a justification for killing (as there might be if this were a case in which killing an innocent person is the only way to save millions of lives). But when Droid engages in this thinking, it does not feel sympathy with the person who might die, and it does not feel anger at the madman. Droid cannot feel guilty if it makes the wrong choice. Is Droid capable of morally evaluating its plans? Can Droid really make a moral judgment? Can Droid act morally?

DOI: 10.4324/9781003402787-8

Historically, emotions have often been thought to be destructive forces, liable to lead us off track, morally speaking. People succumb to appeals to sympathy and betray confidences, fly into jealous rages and kill their spouses, and treat others unfairly out of love for their own kind. Emotions cause trouble and need to be controlled. In this view, emotions are relevant to morality but only because they cause us to act immorally.

If you have some qualms about saying that Droid is truly a moral agent, you will already be thinking that there is something wrong with this historical picture. You would not be alone in thinking this. The view of emotions as an impediment to moral motivation is based on a controversial picture of what emotions are. It sees emotions as blind forces that are not responsive to reasoning, uncontrollable outbursts that can overcome us if we fail to suppress them. Is this what emotions are? Most experts don't think so. Indeed, the most common theories of emotions today recognize that emotions can tell us something about the world (they have intentional content[1]), and that they can be trained and held to standards of appropriateness or justification. As we will see in the first section of this chapter, emotions are quite sophisticated.

If this is true, how should we think about the relationship between morality and emotions? First, emotions might be important causes of morally significant behavior. We have already seen some evidence for this in Part I. The developmentalist psychologists we discussed in Chapter 2 would say that emotions are the natural basis for the moral development needed to act well later in life. Batson's research on the empathy-altruism hypothesis suggests that empathy tends to cause people to perform more altruistic actions.[2] Other psychologists have found that positive emotions promote pro-social behavior such as cooperation and helping (Isen & Levin 1972). The philosopher Myisha Cherry (2021) argues that a particular kind of anger – "Lordean rage," after the Black feminist scholar and poet Audre Lorde – is an important motivation for anti-racist struggle. Lordean rage is directed at those who help to perpetuate racism and it channels the motivational power of anger into action. According to Cherry, there are other types of anger that are destructive, but we are capable of cultivating Lordean rage and doing so will make us better at fighting for justice.

The second potential role for emotions in morality has to do with their connection to moral *judgment*, which will be the subject of much of this chapter. We'll first consider some arguments for thinking that emotions play an essential or constitutive role in moral judgment, a position known as sentimentalism. This discussion will lead us to a distinction in the philosophical literature between moral judgment internalism and moral judgment externalism, which we will consider in the final section of the chapter. Before we get there, let's begin with a survey of different theories of the nature of the emotions.

What Is an Emotion?

We might tend to think of emotions as blind, irrational forces, if we focus on certain emotions such as jealousy or blind rage (not *Lordean* rage.) Often, when we talk about someone who is consumed by jealousy or in a rage, we are talking about someone who has lost control, and these emotions do feel overwhelming and even scary. Think of Bruce Banner, whose anger turns him into a rampaging green monster (The Hulk). But focusing on these examples is misleading. Think instead about the anger you might feel when your college raises tuition rates or the guilt you feel about not visiting your parents at Thanksgiving. Far from being irrational forces, these emotions actually seem to convey something important: in the first case, a conviction about the injustice of the tuition hike, and in the second case, the sense that you ought to have visited your folks. Emotions are *about* something. Anger is directed at those we think have wronged us in some way. We feel guilt about things that we have done that we think are wrong. Pride stems from things we think are good about ourselves, and shame from things we think are bad about ourselves. Fear is a response to perceived danger. These observations about the emotions highlight a key feature of emotions: they are in some way about things we care about or things that matter to us. In the clinical words of psychologist Klaus Scherer, an emotion is a "response to external or internal events of major significance to the organism" (Scherer 2000: 138). The philosopher Peter Goldie calls this the *importance* of emotions (Goldie 2007).

Notice also that we ordinarily think that emotions can be assessed for their *rationality*. We can sensibly ask whether an emotion is appropriate or reasonable. For example, it makes sense to ask whether you should feel guilty about not visiting your parents, or whether it is rational to be afraid of climate change. Most people would say that feeling guilty about not visiting your loving parents is appropriate, but feeling guilty about not buying them an expensive gift is not. This is because we tend to think that people ought to visit their parents at important holidays if they can, but we do not think there's any obligation to buy expensive gifts. Similarly, fear is appropriate when the object of fear is actually dangerous and not when it isn't. Fear of grizzly bears is reasonable if you're hiking in Alaska, but fear of bunny rabbits not so much. Emotions are about something, and they can be evaluated for appropriateness based on their content.

But, of course, emotions also feel like something. Fear makes your heart race, sadness produces a lump in your throat and an inclination to cry, shame makes you turn red and feel hot and like you want to hide. There's even some evidence that the basic emotions feel similar and have the same physiological signs for all people across different cultures.[3] Emotions can also make us do things; that is, they can motivate us to action. You might seek revenge out of anger or try to repair the damage that you've caused out of guilt.

We just ran through five different features of emotions: *intentionality* (they area about something), *importance* (they are evaluative, that is, they are about something that matters), *rationality* (they can be sensibly evaluated for their appropriateness), *phenomenology* (they feel like something), and *motivation* (they tend to cause action). What theorists of emotions in philosophy and psychology are trying to do is to offer theories or models of the emotion that explain these five features. As we'll see in the remainder of this section, different theorists are impressed by (and therefore highlight in their theories) different features. We'll survey three main types of theory: feeling theories, motivational theories, and evaluative theories (of two types: cognitive and perceptual).

In our ordinary way of talking, we tend to use "emotion" and "feeling" more or less interchangeably. If someone asks you how you are feeling, you might very well respond with emotion terms like "happy" or "sad." According to feeling theories, emotions are felt experiences, and this seems like common sense. It's also the view held by many prominent philosophers throughout history, such as Descartes and Hume, and by William James. James' theory (1884) is that an emotion just is the feeling of our body changing in some way. For example, fear is the feeling of our heart rate accelerating, our breathing getting shallower, and our hair standing on end; those feelings constitute the essence of fear. An interesting side note here is that James was a philosopher who was integral to creating the field of psychology – in fact, he is credited with teaching the first psychology class in the United States. Perhaps it is because of this that psychology and philosophy are profoundly integrated in emotions research.

As common sensical as it seems, the feeling theory has some serious problems. The most serious is that thinking of emotions as bodily feelings makes it difficult to understand how emotions are *about* anything. So, feeling theories do very well at capturing *motivation*, but not very well at capturing *intentionality*. When I was sad about getting Covid-19 on my birthday, I was sad *about getting sick on my birthday*. It wasn't just a free-floating bad feeling. Because of this problem, simple feeling theories do not have many defenders these days.

Motivational theories put a different component of emotions at the center: the tendency to cause action (Frijda 1986; Scarantino 2014). According to the simple motivational theory, an emotion is a relatively flexible tendency to cause action. So, for example, fear is constituted by a tendency to take protective actions like running away. Obviously, these theories have an easy time explaining how emotions motivate action! Advocates of the motivational theory also point out that it fits well with the biological function of emotions. Fear likely evolved to make us run away from predators, for example. And motivational theories do make room for the feel of emotions, because it does feel like something to be in a state of readiness to perform an action and these feelings are typically involved in motivating us to do something.

But simple motivational theories have trouble accounting for the way in which emotions are about something and about something that matters. Like feeling theories, they don't do so well at explaining the intentional and evaluative nature of emotions. My tendency to run (or freeze) when I see a grizzly is not *about* the fact that the grizzly is dangerous and this action tendency doesn't represent the grizzly as a threat to my life. Indeed, an action tendency doesn't represent anything at all.

These problems have spurred a new kind of motivational theory, the attitudinal theory, according to which emotions are constituted by the feelings of action tendencies. The main advocates of the attitudinal theory, Julien Deonna and Fabrice Teroni (2014; 2016), hold that action tendencies are attitudes toward the world. For example, the tendency to run away from danger is an attitude toward the threat posed by that danger. The emotion of fear is our feeling of that internal attitude. Deonna and Teroni are trying to build intentionality and evaluation into their theory without abandoning the core idea that emotions are motivational tendencies. They do this by insisting that the attitudes that constitute emotions are not representational attitudes; rather, they are motivational states that convey an evaluation of our circumstances, but not by way of a judgment or belief about those circumstances. This is a subtle point and it's what distinguishing the attitudinal theory from the evaluative theories we will consider next. We could think of it this way: according to the attitudinal theory, an emotion like fear is the feeling of an experience of *grizzly-as-dangerous*. It does not involve a judgment or belief *that* the grizzly is dangerous.

While the addition of evaluative attitudes is an improvement over the simple motivational theory, it's not clear that the way it handles the evaluative aspect of emotions is very plausible. As Christine Tappolet argues, the attitudinal theory seems to get things the wrong way around, because it takes the action tendencies to *be* the evaluations rather than to be responding to evaluations. As Deonna and Teroni put it, "Fear of the dog is an experience of the dog as dangerous, precisely because it consists in feeling the body's readiness to act so as to diminish the dog's likely impact on it (flight, preemptive attack, etc." (Deonna & Teroni 2012: 81). But this doesn't seem right, as Tappolet says: "Do we not aim at diminishing the dog's impact on our body because we consider it to be dangerous? If the answer is yes, it is hard to see how we could do so without representing the dog as dangerous" (Tappolet 2023: 90). In other words, it seems like appreciating the danger is something other than the motivation, which then *causes* us to run, but the attitudinal theory collapses the distinction between the evaluation and the motivation.

Some researchers have been so impressed by the evaluative importance of emotions that they define emotions as evaluations. This is the third type of theory we'll consider. One way to take evaluations to be central to emotions is to adopt a cognitive theory, according to which emotions are evaluative judgments or beliefs. (These theories are also sometimes called "judgment theories.") The Ancient Stoics had a view like this: they thought of emotions as constituted by belief-like judgments. In this view, fear is just the judgment that one is in a dangerous situation and anger is just the judgment that someone has wronged you. Contemporary philosophers such as Solomon (1973) and Nussbaum (2003) have followed suit, though they have developed the view in significant ways. The psychologist Richard Lazarus (1991) also advocated a strong form of cognitivism, according to which "appraisals" (for example, of situations as dangerous) are both necessary and sufficient for an emotion.

Cognitivism has going for it that it has an easy time explaining how emotions are *about* something: they are about the world in the same way beliefs and judgments are about the world. Cognitivism also does well at explaining how we can assess emotions for how reasonable or appropriate they are. This is because if emotions just are judgments or beliefs, then they can be assessed in the same way as judgments or beliefs, that is, according to how accurately they represent what they are supposed to represent about the world. Fear of grizzlies, on this view, is rational because fear is the judgment that grizzlies are dangerous and this judgment is correct.

There is also an old, but entertaining experiment that has been taken to lend empirical support to cognitivism.[4] In this experiment, psychologists Stanley Schachter and Jerome Singer (1962) purported to show that we distinguish our emotions by means of their associated beliefs or judgments. Participants were told that they were helping to determine the effects of vitamin supplements on vision and they consented to being injected with a harmless supplement called "Suproxin." In fact, half the participants were injected with adrenalin and the other half with a placebo. Participants were then put in settings that were meant to induce judgments that are appropriate to either euphoria or anger, two quite different emotions. The setting meant to induce euphoria-appropriate judgments involved a collaborator who ran amok in the waiting room with the subject of the experiment. The description of the behavior of the collaborator makes it sound like a lot of fun to be a psychologist: he wads up paper and plays "basketball" with it, makes paper airplanes, build towers of paper and shoots at them with a rubber-band slingshot, and dances around with a hula hoop. It sounds less fun to be the guy in the setting meant to induce anger-appropriate judgments: he complains a lot and gets indignant about the questionnaire he has been asked to fill out.

Picture the participants in the experiment who got the adrenalin shot. Adrenalin causes rapid breathing and a rise in heart rate and blood pressure, among other things. These participants were either in a room with a guy having a tremendous amount of fun or with a guy getting more and more angry, while their own bodies were exhibiting physical signs of emotion. Schacter and Singer wondered whether the cognitive awareness of the appropriateness of different emotions (euphoria or anger) in the different scenarios would make a difference to what emotion the participants actually took themselves to be experiencing. There was one more important manipulation in the experiment, which was that some subjects were told that "Suproxin" had these side effects (increased heart rate, etc.) and others were not. The thought was that participants who had already had an explanation for their bodily symptoms ("my heart is beating fast because of the drug") would not be influenced by the information from the social setting.

Schachter and Singer found that people in the room with the hula hooper who were ignorant or misinformed about the side effects of Suproxin were more likely to report euphoria and more likely to behave in euphoric ways themselves than subjects who were informed. Uninformed people in the room with the angry guy were more likely to report anger. In other words, the judgment of appropriateness had a strong effect on the emotional experience. Schachter and Singer conclude that "[g]iven a state of physiological arousal for which an individual has no immediate explanation, he will label this state and describe his feelings in terms of the cognitions available to him" (1962: 398). They take this to provide evidence for a cognitive theory of the emotions.

Does this experiment really support the claim that emotions are constituted by cognitive states like judgments or beliefs? Probably not. For one thing, those who favor non-cognitive theories have ways of accommodating the fact that cognitions seem to be necessary for us to label our emotions. And, further, pure cognitivism about the emotions has some problems. For one thing, it takes the feelings out of them, so it has difficulty accounting for the fact that emotions have bodily expressions that are widely shared. For another, cognitive theories have trouble explaining the fact that sometimes our judgments and emotions conflict. It's not impossible to be afraid of something you know isn't dangerous, for example, or to be angry at an inanimate object even though you know it didn't intentionally wrong you. How could this happen if the emotion just is the judgment?

Cognitivism is not the only kind of evaluative theory. There are other ways of understanding the kind of evaluation involved in emotion besides as a judgment or belief, and some of these make more room for feelings. The main alternative we'll discuss is called perceptualism, since it characterizes the evaluations of emotions as akin to perceptions. Emotions, on this view, are like perceptions of color in various ways. Both emotions and color perceptions represent a feature of the world, both have characteristic phenomenology (there is something it's like to experience them), and both can conflict with our judgments (we can see colors that aren't really there just as we can fear something that we don't actually think is dangerous).

Perceptual theories are popular and come in a variety of forms. Some take emotions to be perceptions in a literal sense. According to Jesse Prinz's (2004b; 2007) "embodied appraisal" theory, emotions are perceptions of bodily changes that signify facts about our welfare. So, for example, the feeling of your hair standing on end, heart pounding, rapid breathing, and stomach tightening represents being in danger (which is bad for your welfare) and the perception of this bodily state (the appraisal) is fear. These appraisals still give us information about the world (for example, about whether we are in danger or about to get something really good), but they do so through the body rather than by representing this information in language as the cognitive theory would have it. Prinz's theory has been criticized for identifying the wrong perception with the emotion: if fear is a perception, it sure seems like a perception *of danger*, not of increased heart rate and stomach tightening (Deonna & Teroni 2012).

Other perceptual theories take emotions to be "quasi-perceptions," not of our bodily changes, but of the evaluative features of the objects of the emotions. According to these theories, emotions are not literal perceptual states like sights, sounds, and tastes, but they share features with perceptual states in that they represent features of the world non-conceptually (Goldie 2000; Helm 2001; Tappolet 2020).[5] Basically, emotions, on this view, are felt evaluations. For example, consider again Cherry's Lordean rage, which evaluates the world from the perspective of justice. Lordean rage "represents things like 'racism,' 'racial injustice,' or 'unfairness.' Lordean rage is fitting when it is in response to something racially unjust or insulting – and when something is indeed racially unjust or insulting, it correctly represents [or evaluates] the world" (2021: 38).

The quasi-perceptual evaluative theories have a lot going for them in terms of the five features of emotions identified at the beginning of this section. These theories can explain how emotions are about something (intentionality), because quasi-perceptual states have representational content. They can explain the importance of emotions, because what these quasi-perceptual states represent is evaluative features of their targets. They can explain why emotions feel like something, because quasi-perceptual states – like literal perceptual states – have phenomenological qualities. They can explain how emotions motivate us – certainly better than cognitive theories can – because feelings are motivating. And, finally, they can explain how we evaluate emotions for appropriateness, because we can ask whether the emotion is accurately perceiving the features of the world it targets. For example, Cherry argues that rage is inappropriate when it is directed at scapegoats or at eliminating the other (2021: 37). The idea here is that scapegoats and the mere existence of people who are different from you are not the cause of injustice. So, if your rage is directed at these targets, then it isn't perceiving injustice accurately.

As this brief survey of theories of emotions reveals, emotions are complicated! That seems to be as true in philosophy as it is in life. For our purposes, what's important is what most experts agree on: that a good theory of the emotions must explain how emotions are *about* important features of the world (they are "relevance detectors," as Scherer puts it), and that a good theory must explain why emotions feel like something. It may be that the best theory remains to be discovered. It is also possible that there isn't a single unified theory of emotions and that some emotions are better understood as having a cognitive component, while others are not. In any case, we can safely move forward without deciding which of these views is correct in the knowledge that whatever emotions are, they are not (or not all) blind urges that tell us nothing about the world.

Emotions and Moral Judgment

If emotions tell us something about the world, maybe they tell us something about the moral world. Many of the philosophers we've just discussed think they do; we have already seen how Cherry thinks rage can tell us something about injustice. Indeed, some even think that emotions – as evaluations of one kind or another – are an essential moral capacity, not just for motivating moral behavior but for having any sense of morality at all. That's the idea we'll explore in this section.

Some of my examples above were about anger (or rage) and guilt. Anger and guilt are often taken to be crucial moral motivations because some forms of these emotions seem to be about or directed at moral transgressions. We get angry at people we think have wronged us or someone (or something) we love.

In the Schacter and Singer experiment we discussed, the participants in the anger setting were made to think that they were being disrespected by the experimenters who were wasting their time getting them to answer stupid questions. We feel guilty when we think we've done something bad. If you feel guilty about not visiting your parents at Thanksgiving, it is probably because you have a sense of obligation toward your parents or you think that you ought to do things to make them happy since they went to all that trouble to raise you. Anger and guilt, like other emotions, give us some information about the world, but the information they give us is often particularly morally significant.

There may be other emotions that are morally loaded in this way: for example, empathy, sympathy (Hume's favorite sentiment), moral approval, admiration, pride, shame, resentment, and gratitude all seem to be related to moral judgments in some important way. But in what way? One thought is that emotions are *caused* by moral judgments. When we judge someone to have disrespected or harmed us, we get angry. Emotions, according to this way of thinking, are distinct from moral judgments, though one tends to bring about the other. In another view, emotions are much more intimately involved in moral judgment; they are, indeed, essential to moral judgment. Moral judgments, in this view, just are expressions of our emotions so that judging that someone has acted wrongly is (at least in part) constituted by feeling moral indignation at that person.

In his *Treatise of Human Nature*, David Hume articulates an influential argument for the idea that emotions (or "passions" as he calls them) are essential to moral judgment:

> Since morals, therefore, have an influence on the actions and affections, it follows, that they cannot be deriv'd from reason; and that because reason alone as we have already prov'd, can never have any such influence. Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason.
>
> (2000/1739: 294)

Hume thinks it will be agreed that morality is a "practical" domain (one geared toward action) and that this is confirmed by our experience "which informs us, that men are often govern'd by their duties, and are deter'd from some actions by the opinion of injustice, and impell'd to others by that of obligation" (2000/1739: 457). We tend to be motivated by the moral judgments we make and, furthermore, when our moral judgments change, our motivations tend to change with them. To see this, picture an until now carnivorous friend who decides that eating meat is unethical and starts loudly proclaiming that it's immoral to buy hamburgers. Wouldn't you expect him to at least try to change his behavior eventually? And wouldn't you wonder if he's really sincere about how he feels about the wrongness of eating meat if he went on happily frequenting McDonald's and Burger King forever despite his change of mind?[6]

Moral judgment, in other words, has a special relationship to action that other kinds of judgments do not have. The basic idea is that moral judgments move us in ways that judgments, like "there are three sides to every triangle" or "this book has 12 chapters" do not. Moral judgments must be fundamentally different from these judgments of ordinary fact. Notice that Hume is not saying that moral judgments *always, actually* move us. For one thing, we might not be able to act on our moral judgment. For example, I might think that it's morally required of me to rescue a kitten from a tree, but if there's no ladder and no low branches, there might be nothing I can do. I also might not do anything about the kitten if I'm afraid of heights and my fear overwhelms my moral judgment. Moral judgments *dispose* us to action, then. They tend to make us act, but this tendency can certainly be frustrated by the circumstances or overridden by other motivations. This is the kind of special connection to motivation Hume thinks moral judgments have: they do not always motivate us, but they are essentially tied to motivation. Because Hume thinks judgments made by our reasoning faculty are not essentially motivating (they do not necessarily dispose us to action), he concludes that moral judgments must be made from our sentiments or passions (which *are* motivating, as we discussed in the previous section) rather than from our Reason.

One way to interpret the argument Hume is making here is as a conceptual argument for sentimentalism, which is the view that emotions have an essential or constitutive role in moral judgments. Moral judgments are the kind of thing that motivates people to act and, if that's true, then they must themselves express a motivational state such as a sentiment.

Hume's argument can be buttressed by some empirical findings that show that emotions influence moral judgments. To give you one colorful example, here's an experiment that shows the effect of disgust on moral judgment.[7] Psychologists asked participants to answer questions about the moral propriety of four different scenarios: two having to do with incest between first cousins, one having to do with the decision to drive rather than walk to work, and the last having to do with a studio's decision to release a morally controversial film (Schnall, Haidt, Clore, & Jordan 2008). The participants were divided into three different groups: no-stink, mild-stink and strong-stink. The difference between the three groups was not the stinkiness of the participants, but the amount of stink – in the form of "commercially available fart spray" sprayed into a nearby trash can – in the environment. The results of the experiment were that the feeling of disgust increases people's tendency to make harsh moral judgments. Other experiments have shown that anger makes people more punitive and harsh in their moral judgments about crimes against persons (Lerner, Goldberg, & Tetlock 1998; Seidel & Prinz 2013).

There is also evidence that emotions cause us to make moral judgments that we would not otherwise make. For example, Thalia Wheatley and Jonathan Haidt hypnotized half the participants in one study to feel disgust when they heard the word "often" and the other half to feel disgust when they heard the word "take." All the participants then read some scenarios, one of which was this one:

> Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He [tries to take/often picks] topics that appeal to both professors and students in order to stimulate discussion.
>
> (Wheatley & Haidt 2005: 782)[8]

To those of us who have not been hypnotized, it doesn't seem like Dan has done anything the slightest bit wrong. And, unsurprisingly, participants who read the scenario that did not contain their disgust-inducing word did not rate Dan's behavior as wrong at all. However, for the students who did feel disgust (because they read the scenario with the word that induced disgust for them), there was a tendency to rank Dan's actions as wrong. This is a case in which the people in question would not have made a judgment of moral wrongness at all were it not for the emotion of disgust they experienced.

Now, there is controversy about this evidence. Both philosophers and psychologists have argued that the studies do not really succeed in establishing that disgust has a meaningful influence on moral judgment in general (May 2014; Landy & Goodwin 2015). Moreover, even if the evidence showed what the sentimentalists say it does, the fact that emotions *influence* moral judgments does not establish that moral judgments *are* emotional responses, nor even that emotions are an essential part of moral judgment.[9] This would only be an argument for sentimentalism if the sentimentalist understanding of moral judgment were the *only* way to explain the influence of emotions on moral judgment. Other explanations are possible; it could be that emotions influence moral judgment in the way that wearing rose-colored glasses can influence your judgment about the color of the sky: the glasses influence your judgment, but they're not an inherent part of what it is to make the judgment that the sky is pink. The evidence that emotions cause moral judgments is stronger and more difficult to explain away, but it is still possible for the person who wants to argue against sentimentalism to argue that when moral judgments are entirely caused by emotions they are akin to manipulated illusions; after all, it has not been shown that *all* of our moral judgments are such that we would not make them were it not for our emotions.

The sentimentalist would have a stronger argument if there were empirical evidence that we simply *cannot* make moral judgments without emotions; this sort of connection is what Hume really had in mind. For that we need something more. Some have thought that psychopaths provide some evidence that we can't make moral judgments without emotions, because (to vastly oversimplify) psychopaths are amoral and they do not experience normal emotions like sympathy or compassion. We'll turn to this evidence in the next section, after some stage-setting for another well-known philosophical debate that is at issue here.

Amoralists, Psychopaths, and the Debate between Moral Judgment Internalism and Externalism

In the previous section, we saw that one of the main arguments for thinking that emotions are essentially involved in moral judgment relies on the idea that moral judgments can motivate us directly. Sentimentalism makes moral judgments essentially motivating because emotions, sentiments or passions motivate us. If moral judgments are expressions of sentiments, then it follows that in making a moral judgment we have some motive or other. The claim that moral judgments are essentially motivating is known as *moral judgment internalism*. Sentimentalists are moral judgment internalists (or just internalists, for short). For some philosophers, far from being a point in favor of sentimentalism, the fact that sentimentalism makes moral judgments essentially motivating is a problem for the view. In other words, these people think that *moral judgment internalism* is a mistake. Such people are *moral judgment externalists* (or just externalists, for short).[10]

For centuries, externalist critics of sentimentalism have pointed out that moral judgments can't be essentially motivating, because there are plenty of people who make moral judgments and who aren't motivated by them. Notice that when we talk about moral judgment here, we are not talking about what people *say*, but about the sincere judgments that people form whether they say them out loud or not. Even with this clarification, it's still true that most people have done something they judged to be wrong at the time (at least a little bit wrong), so we are obviously not always motivated to act by our moral judgments. You might think it's quite wrong to lie in a job interview and yet do it out of desperation or because you convince yourself in the moment that "everybody does it." It does seem like a person can make a genuine moral judgment and yet fail to be motivated by it. So, moral judgment internalism has to be formulated in a way that allows for weakness of the will and other kinds of failures of motivation.

The main strategy of response for internalists is to qualify the claim that moral judgments are motivating. Making the internalist claim precise is an interesting philosophical challenge.[11] In general, the claim can be weakened in two ways. First, the sentimentalist can say that the motivation need not be overriding. This is basically Hume's response. Hume responded to this objection by claiming that sometimes the passions that lie behind moral judgments are "calm passions" that do not feel very moving; these passions are moving, but they don't feel the same as, say, lust or greed. To put it simply, you might feel genuinely guilty at the thought of lying on your resume, but your desire for the job overpowers that less-intense moral feeling.

Second, the emotions that bring about moral judgment may be dispositionally motivating, rather than motivating in every single instance. In this view, if you are inclined to judge that killing innocent people is wrong, then you have a background motivating emotion of disapproval of murderers, though it might not motivate you to do anything at the precise moment you make the judgment. When confronted by someone like the amoralist, who makes judgments that use our moral words but is not motivated by these judgments even in the most heavily qualified way, the internalist will say that such a person is not making a *real* moral judgment; rather, they are just pretending or being insincere.[12]

Unfortunately for sentimentalism, these qualifications have not made all the externalist critics convert to internalism. Even with all these qualifications in place, it still seems like there might be people who make moral judgments but are not motivated by them at all in any way. In other words, it seems possible for there to be a true amoralist who sincerely judges (not just in scare quotes) that it is morally wrong to murder someone, for example, but has no disposition to be moved by this whatsoever. Philosophers have traditionally imagined amoralists, but now we have evidence of real people who might fit the bill, namely, psychopaths. Can thinking about psychopaths help us with this debate?

Psychopathy is a personality disorder characterized by impulsivity, egocentrism, lack of empathy, and other traits. The disorder is most often diagnosed by the Psychopathy Checklist, which asks a number of questions that cluster under the headings "aggressive narcissism" and "socially deviant lifestyle" (Hare & Vertommen 2003). Because psychopaths lack empathy, and because they have been thought not to understand morality, they are of interest to those who think emotions like empathy are essential to moral judgment. The basic argument goes this way:

1. Psychopaths do not make a distinction between moral wrongs and conventional wrongs.
2. It is the defect to the emotional response system that is responsible for psychopaths' decreased ability to distinguish moral wrongs from conventional wrongs.[13]
3. Therefore, a functioning emotional response system is essential to moral judgment.

The conclusion of this argument is taken to be strong evidence for sentimentalism (the view that moral judgments express or are about our emotions). Let's look at the steps of this argument in more detail. The first thing to notice is the importance of the distinction between "moral" and "conventional." Conventional norms, such as "you shouldn't go outside in your pajamas," are different from moral norms in a variety of ways. Moral norms are thought to be more serious and have wider applicability than conventional norms. Conventional norms are thought to be contingent on an authority (such as a teacher, the law, or, in the case of the pajamas, a culture), and they receive a different kind of justification from moral norms, which are often justified in terms of harm or fairness (Nichols 2002a; 2004). For example, young children will say that it would be wrong to pull another child's hair, even if the teacher said it was okay, because pulling hair hurts, whereas the wrongness of chewing gum in class depends on the teacher's forbidding it. According to the psychologist Judith Smetana (1981; 1993), this distinction is made from three or four years of age and persists across cultures.

It has been a prominent view that psychopaths don't really understand this distinction (Blair 1995), that is, that psychopaths tend to think of what's morally wrong as what's prohibited by the local authority and they do not see moral transgressions as being more serious than other kinds of violations of rules. This claim about psychopaths in general is now considerably more controversial than it used to be, because of work by Aharoni, Sinnott-Armstrong, and Kiehl (2012). But even this research, which is careful to distinguish different components of what is known as psychopathy, maintains that the "affective defect" part of psychopathy does predict poor performance in distinguishing moral from conventional wrongs. Because they don't feel bad when others suffer, as Jesse Prinz puts it, "they cannot acquire empathetic distress, remorse, or guilt. These emotional deficits seem to be the root cause in their patterns of antisocial behavior" (Prinz 2006: 32). Further, these emotional deficits seem to be responsible for the fact that they don't make the same kind of moral judgments that the rest of us do.

One critic of this argument is Adina Roskies who argues that some people with the same emotional deficits as psychopaths do make real moral judgments. In Roskies' view, patients with *acquired* sociopathy make moral judgments because they "retain the declarative knowledge related to moral issues, and appear to be able to reason morally at a normal level. Significantly, their moral claims accord with those of normals" (Roskies 2003: 57). Acquired sociopaths are people who have the same emotional capacities as psychopaths *now*, but who were normal prior to a brain injury that left them incapacitated. If acquired sociopaths do not have normal emotional responses, but they do make normal moral judgments, then maybe the emotional deficits that psychopaths have are not responsible for their moral problems. It is important to notice that what Roskies means by "normal moral judgment" is different from "moral judgments that conform to the moral/conventional distinction." People with acquired sociopathy have "declarative knowledge" of morality, which means that their moral judgments are about the right things: they'll say that killing innocent people, stealing, lying, and so on are morally wrong. This is not the same as the claim that these people grasp the features of moral judgment that distinguish them from conventional norms (seriousness and authority independence, for example). Roskies' criticism is that some people with emotional defects do make judgments that seem like real moral judgments insofar as they have the right content.

It turns out that looking at psychopaths doesn't really solve the problem by itself. Ultimately, we can see that externalists (like Roskies) and sentimentalist internalists (like Nichols) disagree with each other because they are working with different ideas about what a moral judgment is. Roskies is thinking of moral judgments as defined by their content (whether they accord with judgments of normal people is a matter of whether they are *about* the same things), whereas, for the sentimentalists, moral judgments are defined in terms of the role that they play in our lives (whether they are serious norms that move us independently of the threat of punishment or promise of reward).

Which way should we define *moral judgment*? Is a moral judgment first and foremost a judgment about certain topics (harm, fairness, and so on)? Or is a moral judgment (as Hume thought) an essentially motivating judgment that plays a vital role in our shared lives? We have discussed two types of evidence in favor of sentimentalist internalism (the Humean position): intuitive/conceptual evidence and empirical evidence. Hopes were raised that the empirical evidence would settle things by answering once and for all the question of whether there could be creatures (such as psychopaths) who make moral judgments but are not motivated by them. But now we see that this debate turns on how we understand moral judgment in the first place! Is there anywhere else to look for an answer?

We could take the question of how to define *moral judgment* to be a different kind of empirical question, that is, a question about what people actually mean. To answer it, then, we would need to investigate what people mean when they call something a *moral judgment*. Indeed, Nichols did an empirical study about the normal meaning of *moral judgment* and discovered that people are, in general, willing to say that psychopaths make moral judgments (Nichols 2002a). But what does this show? Nichols himself doesn't think it shows that psychopaths really do make moral judgments. That's because he thinks moral judgment is best understood as a different kind of thing from judgments about social conventions, and people who say that psychopaths do make moral judgments are ignoring this crucial difference. But why put so much weight on the moral/conventional distinction? Why think it tracks what is vitally important to moral judgment? If we find a group of people who claim to make moral judgments and use moral concepts but who treat morals as on par with convention, why say that they're not *really* making moral judgments, rather than just saying that they have a different notion of moral judgment from the rest of us? I think the answer here can only come from thinking more about why we want to know what moral judgments really are and what's at stake in thinking about them in one way or another. In other words, "How should we define *moral judgment*?" is not a purely empirical question; it is a conceptual and deeply theoretical question.

We could advance this theoretical investigation by thinking about why the moral/conventional distinction matters. This distinction seems important to the nature of moral judgment because judgments about violations of serious, authority-independent rules are particularly interesting, both theoretically and practically. They are interesting to philosophers because they raise all sorts of questions about what such judgments could be about. They are practically interesting because they are vitally important to the kind of social regulation that morality is supposed to ensure: it wouldn't be so bad if people started wearing pajamas to work, but it would be a terrible thing if everyone started agreeing with psychopaths about the importance of obeying moral rules. The fact that moral judgments are about serious, authority-independent wrongs is a feature of these judgments that we have particular reason to be interested in, independently of whether this is a feature that belongs to our ordinary concept.

At the same time, "we" (the readers of this book) would not be interested in this feature of moral judgments if it had nothing to do with what people in general mean by *moral judgment*. To answer the question of how we ought to define *moral judgment* we need to think conceptually, theoretically and empirically at the same time. The evidence from psychopathy doesn't prove by itself that emotions are central to moral judgments, but together with a theoretical background that includes an argument for internalism, this evidence adds to the case for a theory of moral judgment that is sentimentalist and internalist.

Taking Stock

We have seen that the best theories of the emotions take emotions to convey (in one way or another) evaluative information about the world. The nature of emotions, then, is at least compatible with the idea that emotions are essential to moral evaluation. We have also seen a number of arguments for sentimentalism, the view that moral judgments do essentially involve or are partly constituted by our emotions: Hume's conceptual argument and several empirical arguments that established links of various strengths between emotions and moral judgments. These arguments can work together to provide a powerful argument for sentimentalism.

If sentimentalism about moral judgment is correct, Droid (the emotionless AI from the start of the chapter) couldn't really be a fully fledged moral agent, because Droid couldn't make real moral judgments. Droid may seem to make moral judgments, but it would really only be faking it. Sentimentalism also entails a compelling thesis about moral motivation, a kind of motivation that Droid would necessarily lack. Intuitively, one thing that seems distinctive about moral motivation is that moral thought – our appreciation of what we have moral reason to do – moves us directly so that as soon as we conceive of something as morally wrong, we are repelled by it, and as soon as we see something as morally admirable, we are drawn to it. Sentimentalism about moral judgment makes sense of this appealing idea, because it takes moral judgments to be constituted by passions that move us.

However important emotions are to moral judgment or moral motivation, we should not forget about moral *reasons*. When we make moral judgments, we tend to think those judgments are based on reasons. For example, your anger at someone who makes a racist joke may alert you to important moral facts, but when you judge that the person was wrong to make the joke, you take it to be based on reasons – of harm or unfair treatment. For this reason, just as we might have qualms about attributing moral agency to Droid, we may also hesitate to attribute moral agency to an animal that has some moral emotions (sympathy, for instance) but no ability to understand or reflect on their moral reasons. The sentimentalist must show that, when we understand moral judgment as a kind of emotional response, we do not lose the connection between moral judgment and reasons altogether. Whether or not sentimentalism can hang on to this connection is a very big question that we will take up in the next chapter.

Summary

- Any good theory of emotions should explain five features of emotions: intentionality (they area about something), importance (they are evaluative, that is, they are about something that *matters*), phenomenology (they feel like something), motivation (a tendency to cause action), and rationality (they can be assessed for appropriateness, fittingness or reasonableness).
- Different theories of the nature of emotions emphasize different features. Feeling theories identify emotions with bodily feelings. Motivational theories define emotions in terms of motivations to action. Evaluative theories (of which we discussed two types, cognitive and perceptual) say that emotions are appraisals or evaluations of the features of their objects.
- Whatever the correct theory of emotions, it is agreed that emotions are not blind urges; rather, emotions can give us information about things that matter to us.
- Sentimentalism in the broadest sense is the view that emotions play an essential or constitutive role in moral judgment.
- There are conceptual arguments for sentimentalism and empirical arguments for the claim that emotions are significantly related to moral judgments.
- Moral judgment internalism is the view that moral judgments are essentially motivating. Moral judgment externalism is the denial of this.
- Sentimentalism is an internalist theory. One counter-example to internalism is the amoralist who makes moral judgments but is not moved by them.
- Some have argued that research on psychopaths (real-life amoralists) can help to settle the debate about moral judgment internalism. But it turns out that whether psychopaths make genuine moral judgments depends on what you mean by *moral judgment* in the first place.
- "What is a moral judgment?" is a question that is partly conceptual, partly theoretical, and partly empirical. Our investigation into the matter should be guided by thinking about the point of moral judgment.

Study Questions

1. Could an artificial intelligence (like Droid discussed in the first paragraph of this chapter) have emotions, according to any of the theories of emotion we discussed?
2. Some Stoics believe that someone who had attained moral perfection would not experience emotions like anger, fear, or guilt. Are they missing something? Are there some emotions we would be better off without?
3. What can we learn from psychopaths that will help us answer our philosophical questions?

4. If you were an exo-anthropologist (in an imagined future in which we can study extra-terrestrial cultures), what criteria would you use to assess whether the aliens are moral agents, or whether their culture has morality? What kind of evidence would you rely on – evidence of their emotional lives, their rational capacities, or both? What evidence would you use to distinguish their morality from their aesthetics or religion?
5. Does Droid make genuine moral judgments? Can Droid do the right thing? Can it act morally? If you think the answers to these questions are different, why?

Notes

[1] The intentional content of an emotion (or a belief) is that toward which it is directed, or what it is about. The word "intentional" can be misleading, if it makes you think of the agent's intentions; that is not what's at issue in "intentional content."

[2] There's an interesting controversy about empathy as a moral motive. Critics of empathy (e.g., Bloom 2017; Prinz 2011) say that it is too partial to be a reliable moral motive. Defenders of empathy (e.g., Kauppinen 2014; Cameron et al. 2022), argue in favor of *regulated* empathy.

[3] Paul Ekman's research on the facial expressions that correspond to different emotions suggests that these expressions are culturally universal (Ekman & Friesen 1971). Not everyone thinks this research really establishes what it claims to. For fascinating reading on cultural diversity and emotions see Lisa Feldman Barrett (2017) and Batja Mesquita (2022).

[4] This experiment has also been thought to provide evidence for social constructionist views about the emotions. See Tappolet (2023) for discussion. Schacter and Singer themselves hold a "two factor" theory of emotions, according to which emotions consist in physiological arousal and cognitive appraisal.

[5] There are different ways of characterizing the quasi-perceptual state. In her "receptive theory," Tappolet (2020; 2023) describes them as "analogue representations."

[6] This is essentially Michael Smith's (1995b: 71–76) argument for what he calls "the practicality requirement."

[7] For more of this story, see Prinz (2007).

[8] Participants were randomly assigned to a group that got one or the other of the phrases in the square brackets; no student saw both the phrases in the brackets.

[9] For an argument against Prinz's use of the empirical studies in particular, see Jones (2006).

[10] Moral judgment internalism and externalism is an entirely different distinction from Reasons existence internalism and externalism. Philosophers apparently love to use the terms *internalism* and *externalism*; there are many more distinctions with these labels in other areas of philosophy! Check the glossary if you become confused.

[11] Almost as difficult as precisely defining the sense in which reasons are internal according to Reasons Internalism! Note again, though, that these are two different kinds of internalism and two different challenges.

12 This is known as the "inverted commas" response (inverted commas are scare quotes), according to which the amoralist who makes moral judgments is making them insincerely, not on her own behalf, but as if she were attributing them to someone else (hence the scare quotes).

13 From now on, for the sake of brevity, I'll talk about the psychopath's *inability* to make this distinction. But it should be noted that what the research shows is that people with the affective defect component of psychopathy are *more likely* to treat moral violations as less serious and more authority dependent than normal people are. It's not true that no psychopath with emotional defects makes any sort of distinction between moral and conventional rules at all. Still, it's the significant difference between psychopaths and normal people that has to be explained. Also, for the sake of brevity, I'll talk about "the psychopath." This is a bit misleading, because in reality people called "psychopaths" are a rather varied group who score higher or lower on different diagnostic criteria for psychopathy.

Further Readings

Aharoni, E., W. Sinnott-Armstrong, & K. A. Kiehl. 2012. Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology 121*(2): 484–497.

Cherry, M. 2021. *The Case for Rage: Why Anger Is Essential to Anti-racist Struggle*. Oxford University Press.

Jones, K. 2006. Metaethics and emotions research: A response to Prinz. *Philosophical Explorations 9*(1): 45–53.

Landy, J. F., & G. P. Goodwin. 2015. Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science 10*(4): 518–536.

May, J. 2014. Does disgust influence moral judgment? *Australasian Journal of Philosophy 92*(1): 125–141.

Nichols, S. 2002a. How psychopaths threaten moral rationalism: Is it irrational to be amoral? *The Monist 85*, 285–304.

Prinz, J. 2006. The emotional basis of moral judgments. *Philosophical Explorations 9*(1): 29–43.

Rosati, C. S. 2016. Moral motivation. *The Stanford Encyclopedia of Philosophy* (Winter edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2016/entries/moral-motivation/

Roskies, A. 2003. Are ethical judgments intrinsically motivational? Lessons from "Acquired Sociopathy." *Philosophical Psychology 16*(1): 51–66.

Tappolet, C. 2023. *Philosophy of Emotion*. Routledge.

7 Sentimentalism and Rationalism

- Rationalism and Sophisticated Sentimentalism
- The Kantian Challenge to Sophisticated Sentimentalism
- The Empirical Threat to Rationalism
- Taking Stock
- Summary
- Study Questions
- Notes
- Further Readings

I am disgusted by people who clip their fingernails in public. I've seen people do it on buses, in airports and even at restaurants, and every time I see it I think, "Gross! Stop doing that. It's disgusting!" But I don't think there's anything wrong with people who are not disgusted by this, nor do I really think that people who do it are bad people. I also think that if I stopped being disgusted by this, I would be just as good a person as I am now. I am also disgusted by the practice of selling children into sexual slavery, which still happens in some parts of the world. But here things are different: I do think there is something wrong with people who are not disgusted by this practice, and I think I would be a much worse person if I stopped being bothered by it. I also think there is something horrifically wrong with the people who treat children this way. It's natural to say that the difference between these two cases is that in the first case I just feel a certain way about a grooming practice that I was raised to think is against etiquette, whereas in the second case I believe that there is a terrible moral violation. If moral judgments are just expressions of or reports about our sentiments, as sentimentalism seems to say, can we really say there is such a difference? Doesn't sentimentalism put my reaction to the nail clipper and my reaction to the child abuser on a par? If so, this would be a serious problem for sentimentalism.

One thing that sentimentalism has going for it is that it helps to explain a distinctive feature of moral motivation, namely, that we seem to be moved to act morally just by the mere thought that "this is the right thing to do." Often when you ask someone why they did something morally good, they will say, "Because it was the right thing to do." The idea here is that in judging that something is the

DOI: 10.4324/9781003402787-9

right thing to do, a good person is motivated in some way to do it (this is, roughly, moral judgment internalism, which we discussed in Chapter 6). You don't need anything more than that. Sentimentalism explains this by identifying the judgment with an emotion that is capable of moving us to action.

In the last chapter, we assumed that the case for moral judgment internalism is good evidence for sentimentalism, because if moral judgments are just expressions of our emotions, then there is an easy explanation for why they motivate us. But, in fact, you don't have to be a sentimentalist to accept moral judgment internalism. It's not only the sentimentalists who claim to explain how we are motivated to act morally in terms of our judgments about the right thing to do. Many rationalists also accept internalism, but reject the crucial role for sentiments in moral judgment.[1] These rationalists think that moral judgments are instead fundamentally tied to our capacity for recognizing rational principles, which can itself motivate us to act.[2]

A rationalist would say that one key difference between my feelings about the nail clipper and my feelings about the child abuser is that in the second case my judgment is based on moral reasons. Furthermore, the rationalist will say that sentimentalism cannot make sense of this difference. Rationalism can make sense of the difference, because it holds that moral judgments are rational judgments that are justified by rational principles, unlike mere tastes (or distastes – as in the case of public nail clipping) that are not underwritten by principles. There are, according to the rationalist, rational principles that determine the truth of our moral judgments. In this chapter we'll consider this objection to sentimentalism and whether rationalism does have a better way of capturing the vast difference between public grooming and child abuse.

Rationalism and Sophisticated Sentimentalism

As I've just hinted at with my opening example, there are some potential problems with the sentimentalist version of internalism (that is, the view that moral judgments are essentially motivating because they are constituted by sentiments) that we have not yet thought about. To see this problem in more detail, we need to think about one way in which moral judgments seem to be different from other kinds of judgments that express feelings. It at least appears that moral judgments, unlike mere judgments of taste, are made for reasons that are supposed to justify these judgments to other people.

Let's start by thinking about, for example, the difference between factual judgments about the health effects of various foods and mere judgments of taste about those foods. Consider the following statements:

- Kale is healthy.
- Kale is disgusting.

If someone told you that you should eat kale because it is healthy, you would expect them to be able to back this up. Even if they didn't have the actual evidence, you would at least expect them to say that they heard on NPR or read in some reliable news source that it reduces the risk of cancer or something like that. If you were to look into it and find the scientist who is making the claim, you would expect the scientist's judgment that kale is healthy to be based on some good reasons that would justify your going out and buying some kale. What about the other judgment? You would expect that the person who says "kale is disgusting" doesn't like kale, but you don't really expect that to be based on any reasons. Indeed, knowing that one person finds kale disgusting doesn't give you much reason to think that you will also find it disgusting, since people have different tastes. The difference here is that "kale is healthy" is a judgment for which there is evidence that is relevant to the rest of us, while "kale is disgusting" expresses a sentiment that there is no reason for the rest of us to share (though some of us might happen to share it). "Kale is healthy" is a claim that makes a demand on the rest of us (that the rest of us believe it) and invites corroboration or debate. "Kale is disgusting" makes no demand on the rest of us to hate kale, and it cannot really be disputed.

The question is, which is moral judgment more like? In thinking about this question, try to put aside any metaphysical worries you might have about realism and relativism, and just consider what you would expect from someone making a moral judgment. Think of one of your friends with whom you have very similar moral beliefs. Let's say you've always agreed with each other that women have a right to abortion. Suddenly, your friend tells you that he has changed his mind. He now thinks that it is morally wrong to have an abortion under any circumstances. Would you expect your friend to give you some reasons for this? Or would you expect him to think of it in the same way as if his feelings about kale had changed: "It's disgusting to me now, but that's just how I feel; I don't have any reason for it." I suspect most of us would be anxious to find out the friend's reasons for changing his mind. We might want to know if these are reasons that we should also be persuaded by, or we might want to know if his values have changed so fundamentally that we must reconsider the friendship. In any case, most of us would expect the friend to have some reasons for the change.

The idea that moral judgments are based on justifying reasons is compelling and fits with our experience. If our moral judgments are supported by reasons, then it might seem like we must – contrary to Hume – reason *to* them using our rational capacities. Just as we reason to the view that kale is healthy on the basis of the evidence about kale and cancer rates, so too we reason to the judgment that abortion is morally permissible or impermissible based on the facts and the moral reasons at stake.

What can a sentimentalist say about these appearances? One thing the sentimentalist might say is that the appearances are misleading.[3] It looks like reasons are involved in our moral judgments, but, in fact, when we seem to give reasons for our moral judgments, what we are really doing is rationalizing them after the fact (*post hoc*). (This is Jonathan Haidt's (2001) view, which we'll talk more about shortly). Maybe we do this because of social pressure to explain ourselves or because of our tendency to look for rational explanations for things, but the truth is (according to this strategy) that moral judgments are not made for reasons in any interesting sense.

If the sentimentalist can't make sense of the apparent connection between moral judgments and reasons in any other way, maybe they will have to accept that there isn't really any connection, but I think the appearances are vivid enough that this strategy should be a last resort. Another strategy the sentimentalist might take starts with the observation that there's a difference between our judgments being *supported by* reasons and our judgments being *the result of* reasoning. The fact that moral judgments are justified by reasons does not necessarily mean that we arrive at our moral judgments by reasoning to them independently of our sentiments – at least not according to *sophisticated sentimentalists*.

In order to make sense of how moral judgments are subject to justification by reasons without abandoning sentimentalism, sophisticated sentimentalists can appeal to the ways in which various sentiments are related to each other. The basic idea, which will take some time to elaborate, is to explain the apparent role of reasons in supporting our moral judgments by complicating the sentimentalist picture so that moral judgments are not simple expressions of our sentiments; instead, they are expressions of sentiments that are the objects of other sentiments we have. When we make a moral judgment of wrongness, in this view, we have sentiments of disapproval toward the wrong action and we also have sentiments of approval toward our sentiments of disapproval. We take a second-order attitude toward our first-order moral sentiments.

According to Allan Gibbard (1992; 2006), these second-order attitudes are endorsements of the appropriateness (or *warrant*, as Gibbard puts it) of guilt and anger, the primary moral emotions. An act is wrong, roughly, if "feelings of resentment or outrage over it are warranted on the part of impartial onlooker and feelings of guilt over it are warranted on the part of the person who does it" (Gibbard 2006: 196). Whether an emotion is warranted or not is to be understood in terms of planning; an emotion is warranted if it is part of the best plan for what it makes sense to do. According to Gibbard, then, when we make a moral judgment we are making a plan for action, where these plans include norms for feelings and actions. For example, if I judge that it is wrong to torment kittens, then I am committing myself to a plan that includes (a) not tormenting any kittens, (b) feeling very guilty if I were to find myself tormenting a kitten, and (c) responding to kitten tormenters with anger.

Gibbard takes these evaluations to be endorsements that are directed at the role of guilt and anger in our plans for action. But there are other forms of sophisticated sentimentalism that describe the second-order attitudes in a slightly different way. For example, a sophisticated sentimentalist could say that the relevant second-order attitudes are approvals of additional norms that make our first-order sentiments non-optional and not dependent on a local authority. In this way of thinking, when I judge that it is morally wrong to torment kittens, I would be feeling a sentiment of disapproval of kitten tormenting in addition to a number of other dispositional attitudes, such as disapproval of anyone who thinks kitten tormenting is only wrong if you think it is or only if local authorities forbid it.[4]

How does sophisticated sentimentalism help us accommodate the idea that moral judgments are made for reasons? It's a complicated story. According to the version of sophisticated sentimentalism we have been considering, the moral judgment "it's wrong to torment kittens" expresses not only a sentiment (of disapproval or anger) toward the action of tormenting kittens, but also some attitudes about this sentiment (such as the attitude of approval toward those who disapprove of this action even when other people are tormenting kittens or the attitude that being angry with kitten tormenters is part of the best plan). Notice first that these second-order attitudes make the judgment in question distinct from judgments like "kale is disgusting." At least for most people who think kale is disgusting, there is no extra disapproval of people who do not find it disgusting, no sense that anger toward people who enjoy kale is warranted. Notice, second, that the way in which moral judgments differ from mere judgments of taste helps us make sense of the idea that we make our moral judgments for reasons.

According to sophisticated sentimentalism, my moral judgment is not just the idiosyncratic expression of a personal taste; rather, because of the way in which it is connected to these other second-order attitudes, it is an expression that makes a claim on others, underwrites judgments about them and may even demand interfering with the behavior of people who aim to do wrong. Therefore, moral judgments are part of a domain of judgments that we take to be justified and that we feel pressure to justify to each other.

Crucially, since sophisticated sentimentalism takes our moral judgments to be linked together in a domain that is subject to pressures of justification, some (sentiment-expressing) judgments can provide reasons for others. For example, if you asked me why I thought it was wrong to torment kittens, I would say such things as "kittens are sentient beings and can suffer" or "people who enjoy gratuitous cruelty are monsters." Here I would be expressing other sentiments that I have about suffering and cruelty. According to sophisticated sentimentalism, it makes sense to say that the fact that kittens would be caused to suffer if we tormented them is our *reason* to be against kitten torment, even though "suffering is bad" and "cruelty is monstrous" are not the deliverances of pure Reason. It makes sense because, according to sophisticated sentimentalism, reasons for moral judgments are also normative claims that must be given a sentimentalist interpretation. Whether one of my judgments counts as a reason for another judgment depends on how exactly they are related – on whether they form a coherent plan or systematic set of attitudes. To be against kitten torment but not think there's anything wrong with causing suffering is less coherent than to be against both. To be against suffering, but not against people who cause suffering is similarly lacking in coherence. Notice that this is not how it is for "kale is disgusting." It isn't incoherent to think kale is gross, but to have nothing against people who love it.

There are deep debates in metaethics about this form of sophisticated sentimentalism (called "expressivism" by its proponents). We can't consider all of them here, but it is worth considering one that has to do with the psychology of moral judgment.[5] Some philosophers have argued on empirical grounds that sophisticated sentimentalism can't be right because it implies that people need to have sophisticated second-order attitudes in order to be able to make moral judgments. Further, they argue, there are obviously people who make moral judgments but who do not have these sophisticated second-order attitudes, such as attitudes about whether guilt and anger are warranted. Shaun Nichols, for example, argues that children who can draw the right distinction between moral and conventional norms are not capable of making normative assessments of the appropriateness of guilt (Nichols 2004: 89–92).[6] To respond to this objection, Gibbard argues that children have "near-moral" concepts that have enough in common with our concepts that we can talk to them about what's wrong, even though they don't have all the capacities they would have to have to make full-blooded moral judgments (Gibbard 2006: 203).

Once again (as in Chapter 6), we can see that two philosophers are disagreeing with each other about moral judgment because they start with different assumptions about the most important features of moral judgment. On one side, the fact that children use moral concepts to make judgments that they think are serious and authority independent is taken to mean that children's moral judgments are real, full-blooded moral judgments. On the other side, the fact that adults make moral judgments for reasons that we offer as justification for the appropriateness of our moral emotions is taken to mean that children probably do not make full-blooded moral judgments. Which side is right? As before when we discussed the different assumptions made about moral judgment by moral judgment internalists and externalists, I don't think there is any way to answer this question without thinking theoretically about what we are interested in and why. For our purposes, it's enough to notice that this debate between Gibbard and Nichols is a "family" dispute among sentimentalists: they both agree that moral judgments and sentiments are intricately intertwined; they just have different views about how to accommodate the idea that our moral judgments are subject to justification by reasons.

Stepping back from the details of various versions of sophisticated sentimentalism, the important point is that these theories provide a critical perspective on our sentiments in one way or another, and this critical perspective allows us to make sense of the idea that moral judgments can be *justified* by reasons. It is worth considering Hume's own view, because it provides a nice illustration of how this sophisticated sentimentalist strategy works. According to Hume, it's not just any old expression of a sentiment that counts as a moral judgment; rather, he says, "Tis only when a character is considered *in general, without reference to our particular interest,* that it causes such a feeling or sentiment as denominates it morally good or evil" (Hume 2000/1739: 303, italics added). Our sentiments constitute moral evaluations when we *correct* them by contemplating the situation from a point of view of sympathy with everyone affected by the action in question. For example, even if you don't feel much of a sentiment about a murder that happened many years ago, you would feel anger about it were you to take the point of view of the people harmed by the crime. Your corrected sentiment of anger toward the murderer constitutes your judgment that the action was wrong. For Hume, then, moral judgments are subject to a standard of correctness given by the proper objects of our moral sentiments, where this standard is characterized by a degree of impartiality and consistency. But the only reason that this standard means anything to us is that we do have a sentiment of sympathy toward our fellow human beings. To express Hume's view as a solution to the problem of how to make room for reasons in sentimentalism, we could say this: The general point of view provides a critical perspective on our moral sentiments, and we can use this general point of view to define reasons that makes some judgments justified and others unjustified. This critical perspective is meaningful to us in a way that influences our moral responses because it engages our sentiment of sympathy.

I've tried to explain how a sophisticated sentimentalist can make sense of one rationalist feature of moral judgment, namely, the fact that we seem to make moral judgments for reasons that we use to justify them to ourselves and each other. A different rationalist claim is that moral requirements are *requirements* of reason in the sense that there are rational principles in virtue of which our moral judgments are correct or incorrect. Rational principles are supposed to be like the principles of logic: true for all time, independent of any empirical facts about us or the world, applicable to all creatures with the capacity to grasp them. Sophisticated sentimentalism does not make this true. There are no rational principles that determine the truth of our moral judgments, according to sophisticated sentimentalism. Though our sentiments are complex and related to each other in various ways that make sense of how they can be justified, it is still sentiments all the way down, according to sophisticated sentimentalism. And according to the rationalists, this is a problem.

The Kantian Challenge to Sophisticated Sentimentalism

Does the sophisticated sentimentalist picture of the relationship between justifying reasons and moral judgments make sense? Does it really capture the sense of normativity we are interested in when we are looking for moral answers? One reason to think that sophisticated sentimentalism is inadequate comes from metaethics: if the reasons that are supposed to justify our moral judgments are just more "endorsements" and expressions of sentiments from us, then they don't really justify anything at all. To go back to our example, the problem is this: how does anything normative come from the added attitude "anger toward the kitten tormenter is appropriate" when this is just another emotional stance that someone has? Closely related to this is a reason to worry from moral psychology: if the reasons that are supposed to justify our moral judgments are just more "endorsements" and expressions of sentiments from us, we aren't (just as a matter of our psychology) going to be able to take them seriously enough.

To understand these problems, it will help to get an idea of what the rationalist alternative is. The rationalist position we will consider here, inspired by Immanuel Kant, is an internalist position (moral judgment internalist, that is), but this internalism takes a different form than it does for sentimentalism. According to the rationalist, the truth of a moral judgment is determined by rational principles; moral judgments are justified, then, insofar as they conform to these principles. That is the sense in which moral judgments are rational, or based in Reason: moral judgments are supposed to tell us what rational principles require of us and they give us reasons to behave in certain ways insofar as they succeed in this aim. In the Kantian picture, people will be motivated by their judgments about what they have moral reason to do *insofar as they are rational*. Moral judgments are essentially motivating, in this view, but only for people whose rational capacities are functioning. A person who thinks that it's wrong to lie in a job interview might nevertheless lie, because she is under so much stress to find a job that some of her rational capacities are overwhelmed in the circumstances.

When we make moral judgments, according to the rationalist, we have reasons for them, just as the sophisticated sentimentalist thinks we do. But for the rationalist the justification of these judgments is not dependent in any way on other sentiments that we have; rather, it is dependent on the authority of certain rational principles. For the sophisticated sentimentalist, ultimately, the explanation for why our second-order attitudes help to justify a judgment makes reference to our sentiments. It's important to keep in mind that the sophisticated sentimentalists do not think that the fact that you have a sentiment against tormenting kittens *is the reason* tormenting kittens is wrong. No: the reason it's wrong to torment kittens is that this hurts kittens. However, according to the sophisticated sentimentalist, if you ask for the ultimate explanation of why we are the kind of creatures who take hurting innocent creatures to be wrong, this explanation will refer to our sentiments. There is no rational principle that *proves* the wrongness. Another way of putting it is this: the reason not to torment kittens is that it causes pain to sentient creatures, but ultimately no consideration would have the *authority* of a normative reason if it weren't for our sentiments.[7]

Kantians think this won't do. It won't do because the authority of our moral reasons will be undermined if it ultimately rests on our sentiments. According to the Kantians, the authority of our moral reasons will also be undermined if it ultimately rests on our desires. Indeed, the Kantian challenge we are discussing here is equally a challenge to those Humeans who favor talking about desires rather than sentiments, though we'll focus on sentiments here.

Christine Korsgaard has made the best case for the Kantian point. Her basic idea is that we are reflective creatures and our reflective nature creates a problem for us, which can only be answered by a Kantian moral theory that allows us to "reflectively endorse" our particular motives and inclinations. "The reflective mind cannot settle for perception and desire, not just as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward" (Korsgaard 1996: 93). According to Korsgaard, moral judgments can only give us reasons if we are able to reflectively endorse them and find them to be completely justified. Our moral judgments must have some justification that bottoms out in rational principles or what Korsgaard calls "laws." If our efforts to justify our moral judgments ended at our sentiments, we would not be satisfied because we can always question whether our sentiments really give us reasons.

What these rational principles or laws are for the Kantian is a long story, which would take us away from moral psychology and into the heart of moral theory. But the basic idea is that rational principles require consistency and that there is a kind of practical consistency that is relevant to morality. The connection between rationality and consistency is a conceptual connection, and it is easy to see in the case of belief. It's paradigmatically irrational to believe a contradiction ("p and not-p"), for instance. How does consistency become relevant to action? For the Kantian there are two ways. First, consistency in action demands that we be able to universalize the principles that we act on. (Kant calls these principles "maxims.") Lying at a job interview in order to get the job is wrong, because in a world in which everyone intended to lie to get ahead, people would expect everyone to lie and lying at job interviews could not work. If you lie at a job interview, you have to be making a special exception for yourself – it's okay for *me* to lie, but I expect most other people not to do it! – and this is a kind of practical inconsistency. Second, Kantians sometimes think of the practical inconsistency in terms of our values. We must value our own rational agency, they say, because without our own capacity for choice we could not value anything at all. But once we see that our own rational agency is the foundation for all the value in our lives, we have to see that it's a pretty special thing and that it is valuable wherever it is. To think that *my* capacity to choose my goals is valuable, but yours is not, also exhibits a kind of practical inconsistency.

So there are two kinds of practical (action-related) inconsistency for Kant: the inconsistency involved in acting for reasons that you expect other people to refrain from acting on (such as "lie to get ahead") and the inconsistency involved in taking your own rational nature to be special, but not according that status to other people's rational nature (which is really just as special). These two kinds of inconsistency are forbidden by the Categorical Imperative, Kant's supreme principle of morality, which tells us only to act on maxims we could universalize and to always value rational nature as an end in itself. We can see intuitively that the Categorical Imperative is a rational law by understanding that it forbids practical inconsistencies.

Excellent moral reasoning for the Kantian, then, will have to include some reflection on the moral law and how it applies to your situation. This will require identifying the maxim of your action (what you're doing and your reason for doing it) and then thinking about what that maxim says about your will and whether it is in accordance with Reason. Does that maxim reveal that you think you're special and deserve better treatment than everyone else? Does that maxim reveal that you have little regard for the value of rational nature? Does your maxim show that you're not following any principle, but just acting on instinct or inclination? If examining what you are doing and why reveals any of these things about yourself, you know that you are on dangerous moral ground.

That was a very condensed explanation of how Kantians think about moral reasoning and the Moral Law, and anyone who is really interested in this should certainly read more about it.[8] But I hope what I've said suffices for us to think about the difference between the Humean sentimentalist and the Kantian rationalist. The difference is in the ultimate explanation they each give for what makes some consideration a justifying reason that has some claim on what we do. For the sentimentalist, at some point the answer is going to be: this is what we care about. For the rationalist, at some point the answer will come to: this is what Reason demands, and anything else is self-contradictory. Korsgaard's point is that the sentimentalist answer is not satisfying because it always makes sense to ask whether we should care about what we happen to care about in a way that it doesn't always make sense to ask whether it should matter if we contradict ourselves.

The Empirical Threat to Rationalism

Many of the sentimentalists whose arguments we considered in Chapter 6 take their arguments for sentimentalism to be arguments against rationalism at the same time. For example, Nichols thinks that the evidence from psychopathy counts against rationalism because psychopaths do not have defects of reasoning and yet do not seem to make moral judgments in the same way that the rest of us do.[9] In this section we will consider the empirical challenge to rationalism by focusing on some work by the psychologist Jonathan Haidt.

Haidt (2001) argues that reasoning does not have the causal role in producing moral judgment that we once thought it had. His well-known article "The Emotional Dog and Its Rational Tail" is structured around four reasons to doubt the causal importance of Reason:

- The dual process problem. We make moral judgments quickly and automatically, and the best explanation of this is that we are using mental processing that is fast, automatic and cognitively undemanding. We use our slow, analytic and controlled mental processing sometimes, but not typically.[10]
- The motivated reasoning problem. Our moral judgments tend to be shaped by a desire to have good relations with other people and a desire to maintain a coherent self-image. Haidt argues that conscious reasoning is more often used to justify the judgments that are motivated by these desires than it is to arrive at the truth.
- The post hoc problem. We can easily construct justifications for intuitive judgments that were not made by reasoning. This causes the illusion of objective reasoning when what is really happening is post hoc (after the fact) rationalization.

- The action problem. Moral action is more strongly correlated with changes in moral emotion than with moral reasoning.

Haidt argues that his own view of moral judgment solves all of these problems. He favors a picture according to which moral judgments are typically made intuitively, on the basis of sentiments (or what he calls "intuitions"). While it is possible for us to reason about our moral judgments, according to Haidt, this happens fairly rarely. He calls his theory of moral judgment "The Social Intuitionist Model" (SIM), because moral judgments are quick, intuitive judgments and, when reasoning is used to make them, it is usually social reasoning that takes place as people talk and argue with each other to try to figure things out. The SIM does allow that individual reasoning or "private reflection" takes place and can have an effect on our judgments, but it is not the usual cause of moral judgment.

We have already seen some evidence that supports Haidt's reasons for doubting the role of Reason. The phenomenon of psychopathy, for example, provides evidence for the action problem, if psychopaths' failure to perform moral actions is explained by an emotional defect. We can't review all of Haidt's evidence here, but there's one piece that has been so widely discussed by philosophers that it's worth looking at in some detail. This is the phenomenon of dumbfounding, which provides some evidence in favor of the post hoc problem.

Moral dumbfounding happens when a person cannot find any reasons for the moral judgment she makes and yet continues to make it anyway. In the widely discussed study that introduced the phenomenon, subjects are presented with the following scenario:

> Julie and Mark, who are brother and sister, are traveling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other. So what do you think about this? Was it wrong for them to have sex?
>
> (Bjorklund, Haidt and Murphy 2000)

Most people say that the siblings' behavior is wrong, and they offer reasons for their judgment. They say Mark and Julie may have a deformed child, that it will ruin their relationship, that it will cause problems in their family, and so on. But because of the way the scenario is constructed, the interviewer can quickly dispel their reasons, which leads to the state of dumbfounding. According to Haidt in an interview about his findings, dumbfounding only bothers certain people:

> For some people it's problematic. They're clearly puzzled, they're clearly reaching, and they seem a little bit flustered. But other people are in a state that Scott Murphy, the honors student who conducted the experiment, calls "comfortably dumbfounded." They say with full poise: "I don't know; I can't explain it; it's just wrong. Period."
>
> (Sommers 2005)

Some people think that the phenomenon of dumbfounding shows that most people don't make moral judgments for reasons.[11] Rather, people offer post hoc rationalizations of their emotional convictions and, when these rationalizations are undermined, they stick with the conviction anyway.

Haidt's research on moral judgment is on the *causes* of moral judgment. Does this research present problems for rationalists? Typically, rationalist moral philosophers such as Korsgaard are not explicitly making claims about the causes of moral judgment, but perhaps rationalists make assumptions that are undermined by Haidt's research. This is the question we will now explore.

First, let's consider whether Haidt and the Kantian rationalists mean the same thing by *reasoning*. If they each mean something different, then Haidt's challenge won't necessarily undermine rationalism. Haidt does seem to have a picture of moral reasoning that is rather different from what rationalists take moral reasoning to be. Haidt talks about the rare cases in which people "reason their way to a judgment by sheer force of logic" (Haidt and Bjorklund 2008: 819). But moral rationalists do not really think that we reason ourselves into moral positions by the sheer force of logic. As we've already touched on, one tool of moral reasoning that Kantians think is particularly important is universalization. In the Kantian picture, when we're unsure what to do, we should ask ourselves whether the intention of our action requires making a special exception for ourselves or whether it is an intention that we think is acceptable for everyone to have. Universalization is like applying the Golden Rule, which tells you to do unto others as you would have them do unto you. (Notice that there is an important difference between the two, however: the Golden Rule asks you to be consistent with how you *want* to be treated, whereas Kantian universalization asks you to think about whether your goals could possibly be achieved in a world in which everyone acted the way you do.) This is a kind of reasoning, one that I would guess is not unfamiliar to readers of this book, but it's not the sheer force of logic.

Still, there is an empirical case against the idea that there is any form of slow, deliberate conscious reasoning in the making of moral judgments, and certainly Kantians think that this kind of reasoning is important. Of course, Kantians do not assume that we engage in this kind of reasoning all the time, nor do they say that our moral judgments are typically *caused* by reasoning. The main point that the Kantian requires is that some moral judgments (the correct ones) are backed up by rational principles and that we *could* – if we needed to – use our rational capacities (such as universalization) to justify these judgments. This doesn't require that we always, or even typically, use our reasoning to arrive at our moral judgments. Indeed it would be a waste of our precious cognitive resources to do this, since most of the moral judgments we make are fairly easy and uncontroversial. When you read in the paper that someone has stolen billions of dollars from the retirement funds of old people or that someone has sold ten-year-old children into slavery, you find yourself making moral judgments about these people. But there's no need for reasoning here; reasoning would be wasted effort since the cases are so obvious. Reasoning is needed in cases of conflict when we aren't sure what to do. For example, what do you do if you rear end someone's car in a parking lot when no one else is looking, causing a small amount of damage? Or if you discover your very good friend cheating on a test? Your automatic judgment might be to do nothing (to drive away, to turn a blind eye for the sake of your friendship), but if you think about it, you might conclude that this isn't really the right thing to do.

Kantians do not need to assume that our moral judgments are always caused by reasoning. What they do assume is that our moral judgments only give us normative reasons if they accurately report what rational principles demand. Moral judgments can be justified and reasoning – when it's done well – produces justification. On the Kantian view, then, it must be that we could reason our way to a moral judgment if we need to, but it's not a problem if many of our actual moral judgments are fairly automatic. Haidt admits that we sometimes arrive at judgments through private reasoning. He also thinks that we engage in social reasoning – reasoning with each other in the form of argument and gossip. Kantians do not need to assume that moral reasoning is always done privately. Indeed, reasoning with each other might help us overcome our biases so that we can be more impartial and better universalizers. Many other empirical approaches to understanding moral judgment also acknowledge a role for reasoning. Shaun Nichols' (2002b; 2004) theory of moral judgment, for example, holds that the causal story of our moral judgments involves two mechanisms: a rational mechanism that has to do with the knowledge of a normative theory prohibiting certain actions, and a sentimental mechanism that generates affective responses to the prohibited actions.

Does the empirical evidence – not just Haidt's work, but all the evidence of the role of sentiments in morality – really provide a fundamental challenge to rationalism? What does seem to be threatened is a picture according to which we always arrive at our moral judgments by engaging in rational reflection on the permissibility of our maxims and we are then motivated to act on these judgments by the sheer recognition of their rational status. It's unlikely that even Kant held this extreme view (Kleingeld 2014). Whether he did or not, it seems to me that the most important Kantian assumptions about reasoning are compatible with much of the empirical research, because Kantians could be satisfied with a limited *causal* role for reasoning. Indeed, Kantians could even admit that emotions have an important role in producing our moral judgments, because this is compatible with thinking that reasoning is how we *justify* our moral judgments and that rational principles are at the foundation of these justifications. As long as we are capable of rejecting an emotionally caused judgment that is found to be unjustified, the Kantian view would not be fundamentally endangered. Furthermore, as long as reasoning can succeed in justifying our moral judgments, it doesn't even have to be the case that reasoning *always* has this purpose. It may be that we often engage in post hoc rationalization in which our aim is just to make ourselves feel better, not to discern any actual rational justification. According to the Kantian, this would be a misuse of our rational capacities, but the fact that these capacities can be abused doesn't mean that they can't also (sometimes) be used well.

Of course, whether reasoning can really provide a justification for our moral judgments depends on some deep issues in metaethics. In particular, it depends on whether there really are any rational principles that provide a foundation for our moral reasons. This is one of the fundamental philosophical disputes between the sentimentalist and the rationalist. If there are no principles of practical reason that have the authority to justify our moral judgments, then Kant was wrong.[12] This debate ultimately depends on a philosophical question about the nature of rationality, not on the psychological facts about the causes of moral judgment. As far as the empirical challenge goes, however, it seems that the door for a modest version of rationalism is still open.

Or is it? There is another problem for rationalism that Haidt's research introduces that we haven't yet considered. Haidt's research seems to cause the most trouble for the rationalist assumption that we are reflective creatures. Recall Korsgaard's claim that "The reflective mind cannot settle for perception and desire, not just as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward" (1996: 93). When our desires or inclinations conflict with each other or with what we deem morally right, we need a conclusive reason to go one way or another, and that reason can't just be another desire or inclination. Do we have reflective minds like this? The phenomenon of dumbfounding might be evidence that we are not reflective in the way Korsgaard thinks we are. Notice that there are really two claims being made here: the first is that we need a reason or a consideration that provides a justification for our action; the second (which depends on the first claim being true) is that this reason cannot ultimately depend on a desire or a sentiment.

The phenomenon of dumbfounding purports to provide evidence that we do not need reasons; some of us do, but others of us are perfectly happy not having any reasons for the moral judgments we make. Notice what a controversial claim this is from the point of view of moral philosophy. We began this chapter with the observation that moral judgments are different from mere judgments of taste insofar as we have reasons for the former but not for the latter. "You morally ought not to eat kale" is importantly different from "kale is disgusting." Sentimentalists and rationalists alike have agreed with this, and sophisticated sentimentalists have bent over backwards trying to accommodate the idea that we (just about all of us) think that our moral judgments should be backed up with reasons. This is a conceptual claim about moral judgments. But if it's really true that nobody except a few philosophers cares about whether they hold their moral judgments for reasons, it does make you wonder whether it really is part of our ordinary concept of a moral judgment that it is supported by reasons. It's worth thinking about what the phenomenon of dumbfounding really shows. To cut to the chase, I don't think the evidence shows that people reject the conception of a moral judgment that philosophers employ (Tiberius 2013).

To have cause to doubt that people care about the reasons for their moral judgments, we would have to see that people

a) do not think there are any reasons for their judgments, and
b) are entirely unperturbed by this fact and have no inclination to reconsider their judgments.

We do not know that (a) is true from the studies that have been done. First, the claim "I don't know why it's wrong, it's just wrong" is ambiguous between "it's wrong for some reason that I don't know" and "there's no reason it's wrong, it just is!" Second, it could also be that people have reasons that they cannot articulate in the moment or do not think count as good enough to offer the person running the experiment. One possibility here that has been studied is that the *risk* of harm – even if no harm occurs – is a reason for judging that an action is wrong (Stanley, Yin, & Sinnott-Armstrong 2019). Consider that we judge people harshly for drunk driving even when they don't happen to get in an accident. So too, people may think it's wrong for Julie and Mark to have sex because there is a real risk of harm that Julie and Mark themselves could not have ruled out.[13]

We also do not know that (b) is true. The fact that people are unwilling to change their judgments in a single interview setting does not mean that they feel no pressure to change them. It may take a long time for shaken confidence to cause someone to change their judgments. Furthermore, given how difficult change is, people may also need some incentive, which they don't really have when it comes to incest. (It would be interesting to know what romantic partners would say if they were convinced that their significant others were actually genetic siblings – would they change their minds about the immorality of incest?) Given this, we would need evidence of how people respond to the challenge to provide justification over the long term. Furthermore, even if a person *never* changes her judgment when she discovers she has no reasons for it, this does not necessarily show that she doesn't care about reasons. There might be other reasons why people fail to change their judgments. For example, maybe there's nothing at stake: I don't know anyone in an incestuous relationship, so there's no practical need for me to rethink the ethics of it. If people do not change their judgments because nothing is at stake or because they just don't feel like thinking about it, this would not count against the claim that people generally take their moral judgments to be justified by reasons. One problem with the study discussed above is that the case presented to the undergraduate student participants (the case of Mark and Julie) did not cause any conflict that mattered to these students. For most North American college students, nothing whatsoever hangs on whether you are for or against incest.

Comfortable dumbfounding is an option here because no justification is practically required.

Of course, the argument I've just made only establishes that it hasn't yet been shown that people don't care about reasons. Notice, though, how unlikely it is that what will be shown is that most people reject a conception of a moral judgment as one that is held for justifying reasons, unlike judgments of mere taste. Some reflection on moral disagreements indicates that people often think of moral judgments as different from judgments of taste. People cast their votes for political candidates who agree with them about the morality of abortion; they march in the streets to protest or support gay marriage; they donate money to charities that help stray animals or feed hungry people. Many people who do these things do them out of moral commitment, and it's hard to imagine that such people don't think there are reasons for the moral judgments upon which they are acting. This is not to say that they're right about this, and it's not to say that there are some cases in which people make moral judgments without much concern for what justifies them. But the idea that moral judgments are different from judgments of taste (like "kale is disgusting") with respect to their being supported by reasons is not just the crazy idea of a few philosophers. The sophisticated sentimentalists are right to bend over backwards to try to explain this.

We have been discussing the first of the Kantian claims listed a few paragraphs earlier, namely, that the reflective mind needs a reason. The second claim is that the reason can't be ultimately explained by a desire or a sentiment. The rationalist thinks that reasons, if they are going to count as genuine reasons that justify what we do, must be explained by rational principles, not desires or sentiments. Their "authority" or justificatory weight cannot ultimately be grounded in how we feel or what we care about. Is this true? This, I think, is the ultimate disagreement between Humean sentimentalists (where we could include sentimentalists and those who think morality is fundamentally about desire) and Kantian rationalists (who think morality is fundamentally about Reason). My own view is that once you see how complex and interrelated our sentiments and desires are, there is nothing threatening in the recognition that they are at the bottom of the explanation of our reasons for action. I think the sophisticated sentimentalists are right. But the Kantians have a point and this is no easy debate. Interestingly, the debate is partly about our psychology and the question, "What are we capable of counting as a justification?" But the debate is also about the normative question of what kinds of considerations actually count as justifying our actions.

There is one more empirical challenge to the Kantian picture, which is a challenge to their particular version of moral judgment internalism. Kantian internalists hold that moral judgments motivate people insofar as they are rational. If moral judgments are as the rationalists think they are, do they motivate us all by themselves? Do our judgments about what rational principles require motivate us insofar as our rational capacities are functioning? If you think, contrary to the Humean Theory of Motivation discussed in Chapter 5, that beliefs can bring about their own motives, then the answer to this question is easy. In this view, beliefs about what rational principles demand can motivate us. But if you were persuaded by the argument in Chapter 5 that beliefs by themselves do not motivate us, then the answer to this question will be more complicated. One way the rationalist might argue that our moral judgments do motivate us insofar as we are rational would be to say that our rational capacities include certain kinds of motives. For example, the rationalist could say that, insofar as we are rational, we have a desire to do what rational principles demand, or she could say that, insofar as we are rational, we have a feeling of respect for rational principles that motivates us to act in accordance with them.[14] This puts Reason in the driver's seat by taking our rational capacities to include capacities to be motivated in certain ways.

On the subject of who is in the driver's seat, we now have more information about what is at stake in this debate. Ultimately, rationalism allows us to hold that our rational capacities are in charge in a very particular way, namely, in a way that is guided by real principles of Reason that justify our actions. The worry about sentimentalism, and sophisticated sentimentalism too, is that if our sentiments are driving, then there really isn't any way to go right or wrong. We drive where we feel like driving, and there isn't really any rule book. The attractive thing about rationalism, for those who are looking for an explanation of the normativity of moral reasons, is that there are rules of the road. I've tried to show that this worry about sentimentalism and sophisticated sentimentalism is unwarranted, because there is a way to go wrong even if our sentiments are in the driver's seat: we can go wrong according to our sentiments, and this is no small thing.

Taking Stock

We began with the idea that reasons are important to moral judgment. This is a point on which sophisticated sentimentalists, rationalists, and many ordinary people agree. The question then became whether sophisticated sentimentalists can make sense of this or whether, if you want to make sense of our moral judgments being backed up by reasons in a way that tastes are not, you have to be a Kantian. In the previous section, I suggested that while the empirical research does not show that rationalism is wrong, it does put some pressure on rationalists to qualify some of their claims. Confronting the empirical evidence about the role of sentiment in moral judgment at least leads us to reject a view of ourselves as primarily rational creatures who always make moral judgments with the aim of figuring out what we really have reason to do and whose rational capacities are always guided by rational principles. We aren't like this. Do Kantians say that we are? Not obviously. Kant himself was well aware of how irrational we can be. Still, once we face the facts about what we are like, we might wonder whether rationalism is really the best theory.

When we consider which theory is best, though, we need to remember that the best theory from the point of view of moral philosophy is one that is compatible with the empirical facts about our psychology *and* able to make sense of how moral judgments sometimes give us justifying reasons for action. In the next chapter we consider a different kind of empirical attack on Kantianism, which claims that it is not rationalist enough.

Summary

- According to one widespread understanding, moral judgments are different from mere judgments of taste insofar as they are justified by reasons.
- Sophisticated sentimentalists try to account for this aspect of moral judgments by showing that our moral sentiments are subject to standards of appropriateness (though these standards ultimately make reference to more sentiments or systems of sentiments, rather than to Reason).
- According to the Kantian rationalist, the truth of a moral judgment is determined by rational principles, and moral judgments are justified insofar as they conform to these principles. People will be motivated by their judgments about what they have moral reason to do insofar as they are rational.
- The rationalist challenge to sophisticated sentimentalism is that it cannot really explain how our moral judgments could ever give us normative reasons, because according to sophisticated sentimentalism the ultimate explanation for the force of these reasons makes reference to our sentiments, and this is not satisfying.
- One empirical challenge to rationalism is that our moral judgments are not caused by our reasoning capacities.
- This empirical challenge misses the mark, because Kantians do not need to assume that moral judgments are typically caused by reasoning as long as (a) we can reason to a moral conclusion when we need to and (b) when we do so correctly we succeed in justifying our moral judgment.
- A different empirical challenge questions the idea that people care about justifying their moral judgments at all. The evidence for this challenge is inconclusive.
- The conception of a moral judgment as supported by reasons is an important one for moral philosophy and it seems unlikely that regular people have no commitment to justifying their moral judgments, for the most part.
- Whether sophisticated sentimentalism is able to explain how moral judgments differ from mere judgments of taste (which are not supported by reasons) or whether rationalism has the better explanation, is at the heart of one of the most important debates in philosophy.

For those of you who find it helpful to see things laid out in a table, I include one here. It refers to a few theories we do not discuss in this book, for the sake of completeness (see Van Roojen (2015) for more on these metaethical positions):

Table 7.1 Moral Judgment and Motivation

	Humean Theory of Motivation • Desires or emotions are necessary to motivate action.	Anti-Humean about Motivation • Desires or emotions are not necessary to motivate action.
Moral Judgment Internalism • Moral judgments essentially motivate the people who make them (at least under certain conditions).	Sentimentalism, Expressivism • Our moral judgments express our feelings or sentiments, which necessarily motivate us.	Rationalism, Kantianism • When our moral judgments are sanctioned by rational principles, they give us reasons that motivate us insofar as we are rational (independently of our sentiments or desires).
Moral Judgment Externalism • Moral judgments do not essentially motivate the people who make them.	Naturalist Moral Realism • Our moral judgments are beliefs about moral facts, which are facts about the natural world. Whether or not these judgments motivate us depends on whether we happen to have the relevant desire.	Non-naturalist Moral Realism • Our moral judgments are beliefs about non-natural (special) moral facts. These beliefs can motivate us by themselves, but they don't necessarily.

Study Questions

1. What do you think is the difference between judgments of taste like "I hate Neapolitan ice cream" and moral judgments like "Slavery is wrong"?
2. How would sentimentalism, sophisticated sentimentalism and rationalism formulate moral judgment internalism? Which is the most plausible formulation?
3. If sentimentalists (like Jonathan Haidt, Shaun Nichols, and Jesse Prinz) are correct about the causes of moral judgment, does this matter for moral philosophy? Does it answer any philosophical questions or rule out any philosophical positions?
4. How do sophisticated sentimentalism and Kantian rationalism attempt to explain the way in which moral judgments give us normative reasons for action?
5. What do you think of Korsgaard's claim that we are reflective creatures of a certain kind? Think of one of your own moral convictions. What discovery would unsettle it? Are there some moral convictions you have that could never be unsettled?

Notes

1 There are also rationalists who are externalists. Their position is of less interest to moral psychology, so we won't be discussing it here.

2 Notice that even if you were convinced in Chapter 5 that desires are necessary for motivation, this rationalist position is still a live option, because the rationalist could say that certain rational states (like respect for the moral law) are sufficiently desire-like to motivate us, or that rationality motivates us by causing new desires.

3 This would have to be the response on behalf of the simplest form of sentimentalism, emotivism: A. J. Ayer's (1952) theory, which has been called the "boo-hurray" theory of moral judgment, because it turns moral judgments like "child abuse is wrong" into "Boo! Child Abuse!" Because of this and other problems for emotivism, it is no longer a serious contender in metaethics.

4 This is basically Simon Blackburn's (1984) position.

5 For more, see Mark Van Roojen's (2015) discussion of non-cognitivism.

6 Moral norms, in distinction from conventional norms, are serious, authority independent and usually justified by appeal to considerations of harm and fairness. See Chapter 6 for discussion.

7 Notice the similarity to what the Humeans say about reasons and desires as discussed in Chapter 5: the *reason* is the fact; the desire is what makes that fact a reason for a particular person.

8 You can start with the original source, Kant's *Groundwork of the Metaphysics of Morals.* Good secondary sources include Baron (1999) and Hill (1992). And for an excellent, accessible guide to living like a Kantian see Stohr (2022).

9 In a very interesting paper called "Do Psychopaths Really Threaten Moral Rationalism?" Jeanette Kennett (2006) argues that they do not (threaten rationalism), because psychopaths also have rational defects. See also Sifferd and Hirstein (2013), whose research on psychopathy will come up in Chapter 11.

10 The idea that we have these two types of mental processing is widely accepted in psychology in some form. For a good resource on what's called *dual-process theory* and these two systems of mental processing, see Frankish (2010). We will discuss dual-process theory again in Chapter 8.

11 It might be more accurate to say the "alleged phenomenon" of dumbfounding. There are many variables in the assessment that someone is dumbfounded that are open to interpretation, for instance, what counts as offering a reason for your moral judgment and what counts as "comfortable" with the inability to justify that judgment. For now, though, let's just accept Haidt's point for the sake of argument and see how far it takes us in the case against rationalism.

12 About metaethics, anyway. One could reject his metaphysical views about rationality but still think he has a lot of important stuff to say about how we ought to treat other people, morally speaking. In other words, one could think Kant had something right in normative ethics even if one doesn't agree with him about metaethics.

13 Another possible explanation is that people think of principles like "incest is wrong" as reasons that support their moral judgments. For relevant research on the effect of principles see Horne, Powell, and Hummel (2015); Lombrozo (2009; and May (2018).

14 The first is, roughly, Smith's (1995b) view; the second is the more traditionally Kantian view.

Further Readings

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Clarendon Press.

Gibbard, A. 2006. Moral feelings and moral concepts. *Oxford Studies in Metaethics 1* (pp. 195–215). Oxford University Press.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review 108*(4): 814–834.

Kennett, J. 2006. Do psychopaths really threaten moral rationalism? *Philosophical Explorations 9*(1): 69–82.

Kleingeld, P. 2014. Debunking confabulation: Emotions and the significance of empirical psychology for Kantian ethics. In *Kant on Emotion and Value* (pp. 146–165). Palgrave Macmillan.

Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge University Press.

Nichols, S. 2002b. Norms with feeling: Towards a psychological account of moral judgment. *Cognition 84*, 221–236.

Stanley, M. L., S. Yin, & W. Sinnott-Armstrong. 2019. A reason-based explanation for moral dumbfounding. *Judgment and Decision Making 14*(2): 120–129.

8 Brains, Biases, and Trolleys

When it comes to many of the examples of morality and immorality we have discussed so far in this book, there's not much doubt about what's right and wrong. As we assumed in previous chapters, it's good to help people and to be kind to your friends. It's wrong to run over someone's dog with your car, to cheat people out of their retirement savings and to molest children. These are not controversial cases. But there are actions that we are much less certain about, and these cases raise the question of how we can know whether something is right or wrong. For example, is it wrong to alter the genes of a human embryo for the purposes of enhancing the resulting person's capacities? Is it wrong to torture someone if you have excellent reason to believe that doing so could prevent the deaths of thousands of people? Many people are also uncertain about how to be moral in their everyday lives: What are the limits of charity and honesty, for instance? Is it enough to buy a gift for Toys for Tots at Christmas, or should you donate 5 percent or even 10 percent of your income to charity? Should you always be honest about painful truths, no matter what the circumstances? If we are uncertain about what the right thing to do is in these cases, how can we try to arrive at an answer?

In general, when we want to figure something out about the world we try to find evidence. More specifically, if we are engaged in scientific inquiry, we might construct a hypothesis about how the world is and then look for evidence that confirms or refutes it. How does this work when the knowledge we are seeking is about what's morally right and wrong? One way it works is pretty simple. You start with a moral theory or principle, and you look for the information you need to apply the principle to your situation. For example,

DOI: 10.4324/9781003402787-10

suppose you start with the utilitarian principle that we should always act to promote the most overall (long-term) happiness for the greatest number of people. With this as your principle, if you're trying to decide whether you ought to tell the truth to your boss about a friendly co-worker stealing from the till, you'll need to know how much happiness and unhappiness will be produced for all concerned, in the long term, by telling the truth and not telling the truth. Suppose you arrive at the conclusion that lying to your boss will produce the most happiness. Is your work done? Not really. There might be other moral reasons to consider, reasons that have to do with fairness or with respecting other people's rational powers. How do you decide which principle is the right one, or which reasons are good reasons to consider? We have quickly arrived at a more fundamental question: how can we know which moral theory or principle is the right one?

Many have thought that at least one crucial kind of evidence for fundamental claims about morality comes from our "intuitions" about what's right and wrong in particular cases. This idea that our intuitions are getting at something important that needs to be taken into account is a central part of the predominant methodology in moral theory: wide reflective equilibrium, which was introduced in Chapter 1. According to wide reflective equilibrium, we evaluate normative theories by bringing into equilibrium ordinary judgments or intuitions about particular cases, putative normative principles, and background (philosophical and scientific) theories (Rawls 1951; Daniels 2008). We may not be able to save all of our intuitive judgments, and some of our principles may need to be modified or thrown out altogether, but the goal is to construct a theory that explains and systematizes as much of this large body of information as possible within the relevant theoretical constraints.

Wide reflective equilibrium raises some questions. What are intuitions? Why do they count as evidence for moral truths and under what conditions are they good evidence? There are many different ways of defining what an intuition is, and many different ideas about why intuitions count as evidence. For our purposes, think of a moral intuition as a moral judgment that appears to be fairly obvious to you without argument or inference. When you see someone setting a cat on fire or kicking a child, it seems wrong to you right away, without having to think about it. This rather immediate judgment is a moral intuition. Psychologists tend to define intuitions in terms of emotional or "gut" responses. Notice that because our definition here uses the term *judgment* (which, as we've seen, can have either a rationalist or a sentimentalist interpretation), it is compatible with thinking that intuitions are more like emotional responses or more like beliefs.

Are moral intuitions good evidence? This is a tough question, and the answer depends in large part on why moral intuitions would count as evidence at all. Instead of trying to answer the question in general, we will consider two specific lines of attack on intuitions and see where they lead us. Both lines of attack draw on research that gives us reasons not to trust our moral intuitions as sources of moral knowledge. We'll review this research first, and then we'll turn to discuss in more detail some of the ways in which philosophers have thought intuitions are a source of moral knowledge. At that point we will be able to evaluate the attack on intuitions.

As we'll see, this attack on intuitions strikes at how we should think about the role of emotions in moral judgment and morality in general. If intuitions are emotional responses to a situation, and if those emotional responses are subject to bias, we may need to think about whether our emotions can always be trusted. Indeed, we've already been introduced to this concern when we discussed research that purports to show that emotions like disgust can influence moral judgments in ways that do not seem morally relevant. Notice that this question about when to trust our emotional responses is important whether you are a sentimentalist or a rationalist, because both admit that emotions influence our moral judgments even if they disagree about the ultimate foundation of morality.

The Attack on Intuitions: Biases and Trolleyology

Thought experiments involving trolleys are now so popular that the research has earned its own nickname: "trolleyology."[1] There are more trolley problem memes than I can count and far too many versions of the trolley case to cover them all. Let's start here: in one study, half of the participants were given this description of the "switch" case:

> *Switch.* A trolley is hurtling down the tracks. There are five innocent people on the track ahead of the trolley, and they will be killed if the trolley continues going straight ahead. There is a spur of track leading off to the side. There is one innocent person on that spur of track. The brakes of the trolley have failed and there is a switch that can be activated to cause the trolley to go to the side track.
>
> You are an innocent bystander (that is, not an employee of the railroad, etc.). You can throw the switch, which will result in the five innocent people on the main track being saved, or you can do nothing, which will result in the one innocent person being saved. What would you do?
>
> (Petrinovich & O'Neill 1996: 149)

Half the subjects in this study were given the scenario as above. The other half had to choose between "throwing the switch, which will result in the death of one innocent person, and doing nothing, which will result in the death of five innocent people." The only difference between the descriptions of the two cases is that one emphasizes the positive side (how many were saved) and the other emphasizes the negative side (how many will die). This difference made a difference: when the positive was emphasized, people were more likely to think that you should pull the switch, whereas when the negative (death) was emphasized, people on average thought you should do nothing.

What's going on here is known as a framing effect (another example of which was discussed in Chapter 1). When the choice is framed positively in terms of how many people will be saved, we tend to think one way. When it's framed negatively in terms of how many will die, we think another way. We tend to put more weight on negative outcomes (people dying), which makes us think that it's worse to let 5 people die than it is to save 1 out of six people. This isn't very logical; it's an example of a kind of bias – negativity bias – that pervades human judgment.[2]

It turns out that negativity bias is not the only kind of bias that influences our moral judgments, and the effects are not limited to thinking about trolleys. In addition, changing the order in which moral cases get presented to people can also change the judgments they make about them. For example, when people were asked about a case in which someone lied and a very similar case in which someone omitted the truth but didn't tell an outright falsehood, how much worse they thought it was to lie outright than to omit the truth depended on the order in which they heard the two cases. Those who heard "omit the truth" first and "lie outright" second were more likely to judge that lying outright is worse than omitting the truth (Haidt & Baron 1996). Surely, though, the order in which you think about two different actions is not relevant to whether those actions are right or wrong. So something is fishy here.

How fishy, though? Those who think we can rely on our intuitions have some responses. First, not many philosophers think that the intuitions we rely on in arriving at moral knowledge are just unfiltered, immediate gut reactions to situations. Instead, they tend to think that the intuitions we should rely on are "considered judgments."[3] That is, they are the judgments we make about situations after some reflection on what's relevant and what is not relevant. Maybe being reflective about our judgments can eliminate some framing effects and biases, particularly because this reflection can include thinking about our biases.

Second, the defender of intuitions will point out that there are certain basic intuitions that are not subject to framing effects. One such defender suggests this as a genuine moral intuition: "The deliberate humiliation, rape, and torture of a child, for no purpose other than the pleasure of the one inflicting such treatment, is immoral" (Shafer-Landau 2008: 83). The critics of intuitions, according to this line of defense, have focused on tricky situations where it's difficult to know what's right or wrong. If we focus on more basic intuitions, we find that there are some that serve as plausible candidates for the foundation of moral knowledge.

At this point in the dialectic, it's worth considering another line of attack on intuitions. This line of attack strikes at the causes of intuitions. The basic idea is that some discoveries about how our moral intuitions are caused should make us doubt their reliability. You can see how this attack might work against trusting our senses about the empirical world. If someone could show that most of what we see is caused by hallucinatory drugs in our drinking water, we would have good reason to think that our visual observations are not a trustworthy source of information about the world. Similarly, if someone could show that our moral intuitions are caused by unreliable mental processes, then we would have some reason not to regard them as a good source of moral knowledge.

Josh Greene, a neuroscientist and philosopher, has made just this argument about a *subset* of our intuitions. Greene argues that different moral intuitions are caused in different ways and that, together with some assumptions about when different mental processes are reliable and when not, we have good reason to discount at least some of our moral intuitions.

To understand Greene's argument, we need more trolleys. Above we considered a case called *Switch*. Now consider two more cases:

> *Footbridge.* A trolley is hurtling down the tracks. There are five innocent people on the track ahead of the trolley, and they will be killed if the trolley continues going straight ahead. You are an innocent bystander (that is, not an employee of the railroad, etc.) standing next to a large man on a footbridge spanning the tracks. The only way to save the five people is to push this man off the footbridge and into the path of the trolley. What would you do?
>
> *Footbridge Switch.* A trolley is hurtling down the tracks. There are five innocent people on the track ahead of the trolley, and they will be killed if the trolley continues going straight ahead. You are an innocent bystander (that is, not an employee of the railroad, etc.) standing next to a switch that opens a trap door, which opens onto the tracks. There is a large man on the trap door. The only way to save the five people is to pull the switch, thus dropping the large man into the path of the trolley. What would you do?

The only difference between *Footbridge* and *Footbridge Switch* is that in the first case you have to push a man to his death with your hands, whereas in the second case you pull a switch that has the same result. Either way, though, the man falls off the footbridge and is killed by the train. Either way, if you act one will die and five will live, and if you don't act five will die and one will live. Despite the similarities between the two cases, people tend to feel very differently about them. Most people say that it's morally permissible to pull the switch in *Footbridge Switch*, but few think that it is morally permissible to push the large man in *Footbridge* (Greene et al. 2001; Greene et al. 2009). This is true not just in western cultures, but all over the world. In a cross-cultural replication of the 2009 study, Bago and colleagues (2020) report that in 45 countries from all the populated continents, people tend to think that personal force (present in *Footbridge* but not in *Footbridge Switch*) makes it wrong to act in a way that would save more lives.

People's feelings follow the same pattern when we compare *Switch* (the first case we talked about) and *Footbridge*: it's okay to pull the switch to divert the train (thus causing the death of the one innocent person who was stuck on the side track), but it's not okay to push the large man for the sake of the same results. Much philosophical ink has been spilt trying to articulate a principle that captures why it is okay to pull the switch but not okay to push the large man. In this effort, it has been taken for granted that our intuitions about the two cases (okay to pull switch/not okay to push man) are onto something and worthy of being accommodated. Hence all the ink: the project of explaining these intuitions by appeal to a principle has seemed like an important project. The introduction of *Footbridge Switch* makes things even more difficult to explain because the cases are even more similar to each other: in both cases you cause a man to be dropped onto the tracks into the path of the train and the only difference is how close you are to this man.

Greene thinks that the difference between *Footbridge* and *Footbridge Switch* can't be explained rationally (Greene et al. 2001). Think about it: in one case you are right next to the man (close enough to touch him), and in the other case you are a little farther away, but you can still make him fall into the train by pulling a switch. How could this tiny difference of physical distance make the difference between its being morally okay to kill him and its being morally wrong to kill him? On the assumption that this tiny difference cannot make a real moral difference, instead of trying to explain our intuitions rationally, Greene sets about trying to explain them causally. His view is that the different intuitions in *Footbridge Switch* and *Footbridge* are explained by the fact that we have two different cognitive systems in our brains. In short, we have one system that is emotional and automatic; this system is engaged when we respond emotionally to the thought of physically touching the man, and it gives rise to the judgment that we should not push the man into the train. The other system is non-emotional and more reflective; when we read the relatively cold *Switch* cases, our emotions are not engaged, so this system can get to work, and it gives rise to the judgment that we should pull the switch in order to save more people. Let's consider this in a little more detail.

The theory that there are these two systems in the brain is called Dual Process Theory (which was briefly introduced in Chapter 7). The first system (System 1) is typically characterized as automatic, emotional, and quick; the second system (System 2) is controlled, deliberate and slow. Greene analogizes the two cognitive systems to the automatic and manual modes on a camera. If you put your camera on automatic, you can take pictures very quickly, but you might sacrifice quality. If you put your camera on manual, you have much more flexibility to cope with different lighting conditions and so on, but you won't be able to take pictures very fast because you have to make a conscious effort to set things up (Greene 2014a).

The two systems are sometimes thought of as "emotional" and "rational", but this is a bit misleading because, in fact, both systems involve affect and cognition.[4] The important point is that the two systems use affect and cognition in a different way. A more precise way of thinking about dual process theory comes from Fiery Cushman (2013) who argues that the crucial difference between the two systems has to do with how the relevant judgments are caused in the brain. System 1 uses "model-free" processing: it assigns a value to an action based on past experience with actions of that type. System 2 uses "model-based" learning: it assigns a value to an action based on a causal model that predicts what the outcomes of this particular action will actually be. As

Greene (2023: 179) puts it, “Model-free learning is a ‘quick and dirty’ way to make decisions with limited computational resources. Model-based learning is the gold standard, provided that one has the resources to execute it well. It’s the difference between relying on a hunch and relying on an explicit understanding of one’s circumstances.”

To get an intuitive grasp of the distinction between “model-free” and “model-based” learning, let me describe the way I navigate the world and the way my sister Paula does. I drive places by listening to the GPS lady tell me which turns to take in which places. After a while, I develop a feel for what I need to do at each intersection, but I have no idea where I am. Paula looks at where she’s going on a map, plans out her route, and knows exactly where she is while she’s driving. After a while, the map is in her head and she can consult that map to know where to turn. Usually, we both get where we’re going, but Paula has the added advantage that if something in the environment changes – say, there’s a detour, or loss of GPS signal – she knows how to accommodate this change, because she can think about alternate routes using her inner map. Because I have no inner map (no “model”), I am stuck turning where I have always turned before or hoping for the GPS to come back online and tell me what to do. When it comes to navigating, I rely more on System 1 and Paula relies more on System 2.

With Dual Process Theory in hand, Greene and colleagues hypothesized an explanation for why people tend to make different judgments in *Switch* and *Footbridge Switch* on the one hand and *Footbridge* on the other. The hypothesis is that our automatic, model-free system of judgment will be triggered by the up close and personal nature of the action we have to perform in *Footbridge* (you have to actually touch the man to push him onto the tracks), and this system will cause us to judge that we should not push the man. On the other hand, in the *Switch* cases, without any emotional trigger, our model-based, calculative system will determine our judgment, and we will consider the outcomes more rationally, thus leading us to say that it would be right to pull the switch. Generalizing, Greene and his colleagues argue that the two processes in Dual Process psychology tend to make different kinds of moral judgments: System 1 produces “characteristically deontological” judgments (judgments naturally justified in terms of rules, rights, and duties); System 2 produces “characteristically consequentialist” judgments (judgments naturally justified by appeal to the greatest benefit for the greatest number).

Psychologists have produced a good deal of evidence in support of this hypothesis. Some of this evidence is neuroscientific: researchers can see from fMRI (functional magnetic resonance imaging) scans that the parts of the brain that are more active when people judge that it would be wrong to push the large man onto the tracks are the parts of the brain that are associated with emotional activity (Shenhav & Greene 2014). More evidence comes from studies of brain-injured patients: in many cases, patients with emotional deficits due to brain injuries are more likely to make consequentialist judgments. (This research gave rise to the fun title: “Consequentialists are Psychopaths”

(Schwitzgebel 2011)). Further, consequentialist judgments are associated with controlled cognition, so that when people are given more time to deliberate or encouraged to reflect, they are more likely to make the consequentialist judgment about a case.[5]

Let's say we accept the description of our psychology put forward by this research: we agree, for the sake of argument, that consequentialist judgments are associated with deliberate and integrative reasoning processes, while deontological judgments are associated with inflexible, automatic cognitive processes. We still haven't reached an illuminating conclusion about *Switch* and *Footbridge*. At this point, Selim Berker (2009), a critic of Greene's work, has argued that there is no bridge from the "is" of Dual Process Theory to the "ought" of ethics. He argues that the scientific evidence about the causes of our moral judgments is irrelevant to any claims about which judgments are right or wrong. According to Berker, to get to any conclusion about which of our intuitions are trustworthy, we would have to rely on moral intuitions about what sorts of features of the world our judgments *ought* to be sensitive to. Only by making such assumptions could we argue that it's *better* to calculate the costs and benefits coldly without being influenced by the "up close and personal" nature of the action.

This is an excellent point: We do need to make some normative assumptions (about what our judgments ought to be sensitive to) in order to get to a normative conclusion (about which judgments we can trust). But Greene does not deny this. Greene's argument is that the scientific evidence *together with* normative assumptions about what counts as good judgment support the conclusion that the consequentialist intuitions are better or more reliable. The argument in favor of trusting consequentialist intuitions depends on the assumption that our judgments *should* (a normative term) not be sensitive to mere personal force. Judgments that respond to these considerations alone – absent any other consideration that could be related to these things, such as special relationships we might have to those who are close to us – are biased by irrelevant information. The scientific research supports the claim that our non-consequentialist judgments really are just responding to personal contact and proximity. The assumption that these features of a situation are irrelevant is an extra moral premise. The extra moral premise, in terms of our example, is that the mere fact that we are farther away from the large man in *Footbridge Switch* than we are in *Footbridge* cannot be morally relevant.

In this section we have seen two reasons to be skeptical about moral intuitions. First, the fact that our intuitions can be biased by irrelevant factors gives us some reason to think that they are not reliable or trustworthy. Second, if a particular subset of our intuitions (the deontological ones) are the result of a cognitive process that isn't designed to respond to the facts of an unfamiliar case (because it is automatic and based only on past experience), then we have some reason to think that these intuitions in particular are not reliable or trustworthy. Having some reason to be skeptical, however, does not mean we should throw out intuitions altogether. We need to evaluate how strong these reasons are and exactly what philosophical conclusions they support. To do that, we need to know more about what role intuitions are supposed to have in the construction of moral knowledge.

Intuitions, Intuitionism, and Reflective Equilibrium

At the beginning of this chapter, I suggested that many philosophers count moral intuitions as evidence for what we should really think about morality. The default method in moral theory, reflective equilibrium, gives intuitions automatic credibility as inputs to our moral deliberations. And even philosophers who do not accept reflective equilibrium as their methodology rely on moral intuitions as evidence.[6] But now we've seen that moral intuitions can be irrational and biased. Why would intuitions be taken as evidence for moral claims? We will consider two basic answers.

First, according to *intuitionism*, moral intuitions have something in common with perception in the realm of scientific discovery. Both intuitions and perceptions purport to track an independent reality. When we see a cat being tortured and we intuit the wrongness of it, we are, in a sense, seeing the wrongness. Intuitions are good evidence insofar as they really track the moral truths they claim to track, just as our visual perception is good evidence about the physical world as long as we are perceiving accurately. Historically, some intuitionists have thought that intuition is a special faculty – something like a sixth sense – that perceives moral truths. Recently, this view has fallen out of favor because the "special faculty" seems very mysterious. Contemporary intuitionists tend to think that our moral intuitions are just beliefs that do not require any special faculty beyond the rational and perceptual capacities we already have. In the previous section we saw an example from Russ Shafer-Landau (the most prominent contemporary defender of intuitionism) of a reliable moral intuition: "The deliberate humiliation, rape, and torture of a child, for no purpose other than the pleasure of the one inflicting such treatment, is immoral" (2008: 83). The important thing about this belief, according to the intuitionists, is that it does not need to be inferred or deduced from other beliefs to be justified; you know it is true just by understanding it. The intuition Shafer-Landau describes is *self-justifying*, and this is why it is supposed to constitute evidence relevant to our moral knowledge. Self-justifying beliefs are a secure foundation for moral knowledge.

Second, some philosophers think that moral intuitions are not evidence of an independent moral reality or moral properties that exist outside of us, yet they are nevertheless indispensable starting points. In this way of thinking, moral intuitions (or considered judgments), are the building blocks with which we cannot but attempt to construct a moral system that works for us. We cannot avoid starting with our own moral intuitions because our goal is to improve what we think for the purpose of getting along in life, and our moral intuitions are just what we think. Moral intuitions are the convictions that it is the business of moral thinking to evaluate. This interpretation of moral inquiry makes the project of finding moral knowledge a project of construction. Moral knowledge is constructed by a method that takes our intuitions or considered judgments about morality and refines them through a reflective process. Hence, this way of thinking about morality is called *constructivism* (Rawls 1980).

Does the psychological research we have been discussing cause problems for intuitionism? If our moral intuitions are systematically biased, then it looks like if we do have a special faculty for discerning a realm of moral truths, it isn't very reliable – it's as if we had evidence that we frequently "see" things that aren't really there. Furthermore, if different people have different intuitions about the same case and it turns out that the explanation for this is that the two groups of people are using different cognitive processes, then again it looks like we don't have a faculty of moral intuition that gives us access to the untainted moral facts. Rather, it looks like we have different faculties, designed for solving different kinds of problems, none of which is the problem of perceiving an independent realm of moral truths.

Of course, contemporary intuitionists do not think we have a special faculty of moral intuition. Rather, they think we discern moral truths using the ordinary mental capacities psychologists agree we have. Does the evidence against intuitions cause problems for this view? If we think again about our example (the intuition about the wrongness of torturing a child for fun), the research we have looked at doesn't give us any reason to doubt *this* intuition. How might the critic of intuitions press her case? First, she could say that genuinely self-evident intuitions are so specific that they can't help us in hard cases, nor do they provide a sufficient foundation for a moral theory. In other words, the critic could say: yes, *that* intuition is reliable, but it doesn't give us enough knowledge to get us to a complete moral theory. Second, the critic could argue that the fact that there is an intuition we have no reason to doubt does not mean that our intuitions are generally reliable and that they should get automatic credence in our moral deliberations. If there is a good argument for thinking that whatever faculties produce our moral intuitions often go awry, it doesn't help if we can find one case where they probably didn't. After all, that one case could just be luck. The problem with luck is that when we confront cases about which we're unsure what it's morally right to do, we won't know if our intuition about that case is trustworthy or not (we won't know if we happened to get lucky). Sure, we know we shouldn't torture children for fun, but what about incest or pushing large men into trolleys?

The contemporary intuitionist might respond to the second criticism that the question of whether a particular moral intuition is reliable or not is answered through the process of reflective equilibrium. Through this process we can think about whether some of our intuitions were produced in ways that make them defective. For example, we can think about whether an intuition was the result of bias, and we can question what it means if it conflicts with another intuition. Just as we might question some of our visual perceptions when we are trying to acquire knowledge of the physical world – "Was the light good?", "Were my eyes tired?" – so too we can ask whether our intuitions are credible moral intuitions on the basis of other knowledge we have.

Constructivists also use reflective equilibrium as a method for acquiring moral knowledge. Constructivists don't think that moral intuitions are like perceptions of an independent reality, but they do think they are the building blocks of moral knowledge. Constructivists disagree with intuitionists, then, about *why* we should take account of moral intuitions, but they both agree that they have to be taken seriously. For the intuitionists, moral intuitions must be inputs to the search for moral knowledge because they are our access to the independent moral reality. For constructivists, moral intuitions must be inputs to the search for moral knowledge because the point of moral inquiry is to refine the convictions that we start with in a way that can help us get along and thrive. Constructivists don't have to worry in the same way as intuitionists do about the possibility that our mental faculties are not reliably tracking an independent moral reality (because they don't think there is such a thing), but they do need to worry about the reliability of reflective equilibrium, the main method of moral inquiry.

So the important question is this: Does the psychological research we have considered cause problems for reflective equilibrium? I think the answer is: not really. It might, in fact, *inform* reflective equilibrium in an important way. After all, reflective equilibrium requires us to bring together into a coherent whole our considered judgments, our ethical principles *and* any relevant background scientific and philosophical theories. In this way of thinking, psychological theories about biases that affect judgment, or about dual cognitive processes, should be brought into equilibrium with everything else.[7]

Once we see things this way, we can recognize that "trolleyology" is really about figuring out what we should think. Should we think it's wrong to push a large man into an oncoming train if doing so will save five people? Yes, this is how most of us tend to feel, but is it what we *should* think after we've reflectively considered all the angles? Should we think that it matters whether you are pushing a large man or flipping a switch when the same number of people will live and die in either case? Certainly, there doesn't seem to be a huge moral difference between flipping a switch that will kill someone and pushing someone to his death. As we have seen, Greene argues partly on this basis that the non-consequentialist intuition isn't to be trusted in the footbridge case. But there is another way to go.

Judith Jarvis Thomson agrees that there is an uncomfortable incongruity between our willingness to flip the switch and our unwillingness to push the large man. We don't have much to draw on to justify this pair of judgments. "Well, in *Switch* I don't have to use both hands!" doesn't seem like much of a moral argument. But instead of concluding that we ought to count both pulling the switch and pushing the large man as the right thing to do (as Greene does), Thomson (2008) argues that we ought to judge both actions to be wrong. She does this by introducing another case in which *you* are the person on the side track and if you pull the switch you would kill yourself, thereby saving the five innocent people. Sure, it would be nice if you did this, but, Thomson argues, it is not morally required of you to sacrifice yourself to save the five. Sacrificing yourself to save five others would be heroic or supererogatory (beyond the call of duty), but it isn't required. Further, if you aren't required to sacrifice yourself, then the stranger on the track in *Switch* isn't required to sacrifice himself to save five either. By switching the train onto his track, you make him do something that he isn't required to do and something that you yourself would not do. This seems wrong. This new twist on the case makes us think that maybe we ought to make the same judgment in *Switch* and *Footbridge*: in neither case does morality require you to do what you need to do to save the five.

This conclusion is even more anti-consequentialist than the usual intuitions people have about these cases – it says that we should pay even less attention to the cost-benefit analysis of how many would live and how many would die than we originally thought – and it would probably make Greene's head spin.[8] But it is an option. It is also a very interesting illustration about how reflective equilibrium can work. Thomson originally thought, with most people, that we should pull the switch in *Switch* but we should not push the large man in *Footbridge*. But then she changed her mind, in part because she was persuaded that there was no principle that could explain why it's right to pull the switch but wrong to push the man (Thomson 2008). What happened here is that the attempt to reach reflective equilibrium caused her to have to jettison her original intuition about pulling the switch. In contrast, because of the assumption he makes about the reliability of System 1 judgments in novel cases, Greene jettisons the intuition that we should not push the large man.[9]

There is another vein of psychological research that bears on this debate. Karen Huang and her colleagues have been investigating how veil of ignorance reasoning affects people's judgments about trolley cases, and finding some interesting results. The veil of ignorance was introduced by John Rawls (1971) as a heuristic for securing fair, impartial reasoning. The idea is that when you're thinking about a problem – Rawls's interest was in how to distribute scarce resources – you should abstract away from the specifics about yourself (your race, gender, religion, wealth, and so on), as if you were standing behind a veil of ignorance deciding what to do before you know what position you will occupy. From this position, you make the decision that is best no matter who you turn out to be in life. Rawls thought that this procedure was the right way to think about principles of justice, but it can also be used to think about trolley scenarios. The idea here is that you think about the right thing to do not knowing if you will be the one person who would be killed (either by the diverted train or the push) or among the unlucky five at whom the train is pointed. As it turns out, when people are encouraged to think about the trolley problems using veil of ignorance reasoning, they are more likely to think that whatever is necessary (including pushing the large man) should be done to save the five (Huang, Greene, & Bazerman 2019). This makes sense – because it's five times more likely that you'll be one of the five! This research puts some pressure on the anti-utilitarian solution to the trolley problem, because it provides evidence that impartial reasoning favors thinking in more consequentialist terms.

How will this disagreement be resolved? Unfortunately, reflective equilibrium doesn't give us an easy way out. There's no answer sheet and no simple formula for making the right decision. But, then, no important human inquiry is like this. Instead, we have to engage in the messy process of evaluating all the pieces and trying to put them together into something that makes sense. I hope this section has at least added to the case that the facts about our psychology are relevant to this process.

Taking Stock

This chapter has focused mainly on the relationship between empirical research on the causes of moral judgment and questions about the role of intuitive judgments in moral epistemology. We've seen that moral judgments may be pushed around by biases or caused by unreliable processes, and if they are, then we should take this into account when we are engaged in moral reflection. Along the way, there have been many hints that these topics bear on other questions in moral psychology. It's worth closing with a few observations about the relevance of trolleyology to other debates.

First, consider the debate between sentimentalism and rationalism discussed in the previous chapter. Notice that the psychologists and neuroscientists whose research we have discussed in this chapter all accept that moral judgments are a product of both sentiment and reasoning. Greene argues that automatic emotional responses can sometimes lead us astray and that we are well advised to temper our intuitions with reason, but he doesn't think that emotions always lead us astray. Intuitive, emotional reactions against killing other people for the most part lead us down the right path. So, the trolleyology research assumes a sophisticated answer to the emotion vs. reason debate about moral judgment, one which is likely to be compatible with sophisticated versions of both sentimentalism and rationalism.

Second, consider the debate about which normative moral theory is correct, consequentialism (utilitarianism) or deontology (Kantianism). To my mind, both of these normative theories have been caricatured in the trolley wars. Consequentialism looks like a theory fit for psychopaths whose main purpose is to push people into oncoming trains. Kantianism looks like a theory for the most rigid rule followers who understand nothing about human nature. Neither of these caricatures stands up to the nuanced developments of these theories in the literature. Some consequentialists think that the best version of the theory is indirect, and no consequentialist seriously recommends that we go around constantly calculating costs and benefits (Hooker 2002; Railton 1984). Some Kantians argue that it does sometimes make sense to think about the numbers of lives that would be lost if we refrained from taking action (Hill 1992) and some argue that Kantians should pay close attention to the empirical literature in moral psychology (Kleingeld 2014). Empirical psychology is relevant to moral philosophy, but care needs to be taken in applying one to the other.

Summary

- Some moral questions (like whether it's wrong to cheat people out of their retirement savings) are fairly easy, but we do sometimes face moral questions to which we don't know the answers. When this happens, we seek moral knowledge, or at least reasonable moral judgments.
- A moral intuition is a moral judgment that seems true without having to engage in inferential reasoning. Intuitions about cases are often thought to be evidence for moral conclusions.
- Recent studies about the psychology of judgment give us reasons to be skeptical about the value of our intuitions for arriving at moral knowledge.
- One source of skepticism is the fact that our judgments are subject to biases, such as negativity bias.
- Another source of skepticism is the conflicting intuitions we have about trolley cases that differ in morally insignificant ways.
- Dual Process Theory, which says that our moral intuitions are the result of different cognitive systems, is one explanation for why we have these conflicting intuitions. Joshua Greene argues that our quick System 1 processing is not trustworthy in new situations because it is an automatic system that doesn't pause to consider the novel circumstances.
- Intuitionists and constructivists agree that moral intuitions must be taken into account in the search for moral knowledge, though for different reasons.
- The psychological research does not cause problems for reflective equilibrium, a method that could be used by both intuitionists and constructivists. Indeed, psychological research can inform reflective equilibrium by showing us the conditions under which our automatic responses are not necessarily to be trusted.

Study Questions

1. Some have complained that trolley cases are too far from real life for us to learn much from them. Are there more "real life" cases that have the same features as Switch, Footbridge, and Footbridge Switch?
2. Think of some moral intuitions that you have. What do you think would be required for you to make these into "considered judgments"? What sort of standards should we apply to our intuitions if we are to trust them?
3. Greene and Thomson resolve the Switch/Footbridge quandary in opposite ways. What do you think is the right solution?
4. Are there cases in which System 1 moral thinking leads us in the right direction? What would happen if we didn't have it?
5. Think of a moral quandary that you have experienced. How would you proceed to figure out what to do using wide reflective equilibrium? Do you see any pitfalls in this procedure?

Notes

1 Philippa Foot (2002/1967) originally introduced runaway trolleys (or "trams" as she called them) as part of an argument for the doctrine of double effect, which is the theory that it is permissible to do something that will result in a morally bad outcome if that outcome is a side-effect of what you intend rather than intended. Judith Jarvis Thomson (1976) is the other key player in the history of trolleys.

2 For an excellent discussion of this research, see Sinnott-Armstrong (2008).

3 This is what John Rawls (1971), whose worked helped make reflective equilibrium the default method in moral theory, calls the intuitions we should attend to. Intuitions understood this way require clear thinking about the case, though they are still intuitions by our definition because we do not arrive at them by way of argument or inference from a principle.

4 It may even be misleading to think of the two systems as "fast" and "slow" (Bago & DeNeys 2017).

5 See Greene (2014a) for an overview of these studies that includes more helpful references.

6 Antti Kauppinen (2014b) argues that reliance on intuitions may be inevitable in moral theory, and he takes that to be a problem for Greene.

7 Greene (2014a: 726) agrees: "Along with our 'considered judgments' and organizing principles, we must add to the mix a scientific understanding of the psychological and biological processes that have produced them. (Call this *double-wide* reflective equilibrium)."

8 See also John Taurek's (1977) rejection of cost-benefit analysis when it comes to the moral rights of individuals.

9 Things are not quite this simple, of course. Greene is also motivated by the fact that Thomson's argument puts a lot of weight on the distinction between doing something and allowing something to happen. After all, her argument leads to the conclusion that we may not pull the switch if that will kill one person to save five. Greene thinks this gives far too much important to the arbitrary fact of what position the switch was left in.

Further Readings

Bago, B., B. Aczel, Z. Kekecs, J. Protzko, M. Kovacs, T. Nagy, T. Gill, U.-D. Reips … & C. R. Chartier. 2020. (Preregistered and conditionally accepted). Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgments. *Nature Human Behaviour*.

Berker, S. 2009. The normative insignificance of neuroscience. *Philosophy & Public Affairs 37*(4): 293–329.

Foot, P. 2002/1967. The problem of abortion and the doctrine of double effect. Reprinted in *Virtues and Vices and Other Essays in Moral Philosophy*, pp. 19–32. Oxford University Press.

Greene, J. 2014a. Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics 124*(4): 695–726.

Greene, J., F. L. Cushman, K. Stewart, L. Lowenberg, L. Nystrom, & J. Cohen. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition 111*(3): 364–371.

Huang, K., J. D. Greene, & M. Bazerman. 2019. Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences 116*(48): 23989–23995.

Kauppinen, A. 2014b. Ethics and empirical psychology: Critical remarks to empirically informed ethics. In M. E. Christen, C. E. van Schaik, J. E. Fischer, M. E. Huppenbauer, & C. E. Tanner (eds), *Empirically Informed Ethics: Morality between Facts and Norms* (pp. 279–305). Springer International.

Kleingeld, P. 2014. Debunking confabulation: Emotions and the significance of empirical psychology for Kantian ethics. In *Kant on Emotion and Value*, pp. 146–165. Palgrave Macmillan.

Lillehammer, H. 2023. *The Trolley Problem*. Cambridge University Press.

Shafer-Landau, R. 2008. Defending ethical intuitionism. In *Moral Psychology*, Vol. *2: The Cognitive Science of Morality: Intuitions and Diversity*, W. Sinnott-Armstrong (ed.), pp. 83–96. MIT Press.

Sinnott-Armstrong, W. 2008. Framing moral intuitions. In *Moral Psychology,* Vol. *2: The Cognitive Science of Morality: Intuitions and Diversity*, W. Sinnott-Armstrong (ed.), pp. 47–76. MIT Press.

Thomson, J. J. 2008. Turning the trolley. *Philosophy & Public Affairs 36*(4): 359–374.

9 Virtue

- What Kind of State Is a Virtue?
- Should We Aspire to Virtue? The Empirical Challenge
- Defending Virtue
- Taking Stock
- Summary
- Study Questions
- Notes
- Further Readings

Why did you take in the stray cat? Or help your neighbor move into his apartment? Why do you recycle? So far we have considered two basic possibilities: you do these things because you want to, or you do them because you think you should (where "thinking that you should" could be understood in sentimentalist or rationalist terms). But maybe you took in the cat or helped your neighbor because you're a kind person. Maybe you recycle because you're civic-minded. In other words, maybe it's really our character that explains why we act morally when we do. As it turns out, this is an ancient idea.

It's a very attractive idea from the point of view of trying to capture what is special about moral motivation. The desire view doesn't obviously do a good job of this. If we act morally simply because we want to, this doesn't make moral motivation distinctive in any way: we do everything because we want to! The desire theory at least owes us an explanation of which desires are morally special. Kantian rationalism does a better job of making moral motivation special, particularly if we believe Kant that moral motivation is the motive of duty. But as we saw in Chapter 5, the pure sense of duty didn't seem like it could possibly be all there is to moral motivation. Virtue, on the other hand, makes moral motivation distinctive, and it also allows for many different types of moral motives – as many as there are virtues.

Not only does virtue provide a distinctive form of moral motivation, it also underlies a very familiar form of evaluation that we engage in all the time: the evaluation of people in terms of their character. To see what I mean about this being familiar, think about the kinds of things we say about political candidates.

DOI: 10.4324/9781003402787-11

We tend to talk about whether they are honest and fair, whether they have integrity or the courage of their convictions, and sometimes we just flat out ask whether they have good character. Character judgments have even made it to bumper stickers, as in the once popular "Mean People Suck." Assessments of character also pervade our personal lives. We gossip about whether someone is really arrogant or just insecure, whether someone else is overly trusting or gullible. We praise some friends for their generosity and others for their honesty. When friends ask us our opinions about their dating partners, we tend to consider character: Is he really as nice as he seems? Isn't that one kind of manipulative? It's hard to deny that personality matters to us; no one wants to date a jerk. Evaluations of actions in terms of the character of the person performing the action are also familiar: helping that elderly person across the street was a kind thing to do, rescuing the child from the burning building was brave, and so on.

Virtue ethics is the name for a family of ethical theories that place the virtues at the center of ethics (Hursthouse & Pettigrove 2022). Given how much we value character in our ordinary lives, virtue ethics has a lot going for it. As we'll see, however, there is also reason to worry about putting too much emphasis on virtue.

What Kind of State Is a Virtue?

According to Aristotle, who is the inspiration for much modern virtue ethics, a virtue is a tendency or disposition, cultivated by habit, to have appropriate beliefs and feelings and to act accordingly (*Nicomachean Ethics*: 1105b25–6). The fully virtuous person feels anger, pity and other feelings "at the right times, about the right things, towards the right people, for the right end, and in the right way" (*Nic. Ethics*: 1106b21ff). Aristotle thought that the right time, target, and so on, for most virtues is at the mean between two extremes. Courage, for example, is the state of character in which one has appropriate fear: too much fear makes you a coward and too little fear makes you rash. Temperance is the mean between insensibility (an inability to appreciate pleasure) and intemperance (the tendency to sacrifice long-term benefits for pleasure in the short term). Finally, virtues are supposed to be deeply entrenched in the person who has them, not fickle or wavering. A person who has the virtues will tend to feel, think, and act correctly in different situations and at various times.

Aristotle thought that virtues are the traits that are essential to human flourishing and he thought that we should understand what a good or flourishing life is for a human being by understanding the function of a human being. The general thought here is that judgments of the form X is *good for* Y are always made relative to the kind of thing that Y is and what it is to be a good one of those. A good knife is a knife that cuts well, and so what is good for a knife is to be kept sharp. A good bee is one that performs its function in the hive, so what is good for a worker bee is whatever enables him to find pollen, and what is good for the queen bee is whatever enables her to produce lots more bees. A good lioness is one that can hunt down her prey and feed her cubs, so what's good for a lioness is to have sharp claws and powerful legs. So too what is good for a human being is to be good at whatever it is human beings are supposed to do, that is, what is good for humans is to fulfill our natural telos.

The special nature of a human being, according to Aristotle, is that we are beings who can guide our actions by using our capacity to reason. We are also physical beings for whom social interaction with other human beings is important. The human telos, or function, then, is to use reason to think and feel appropriately in all of our endeavors, but especially in our interactions with others. To perform this function well is to do it excellently, or in accordance with virtue. According to Aristotle, we flourish when we exercise the virtues that make us the best exemplars of our rational, emotional, social kind. In short, Aristotle's view is that because of the kind of beings we are, we will only flourish if we are virtuous: courageous, temperate, just, generous, wise, and so forth. You can see how these traditional virtues are good for human beings if you think about the kinds of things human beings typically do. We set goals for ourselves (such as graduating from college) that require making short-term sacrifices for long-term benefits, and this requires temperance. In pursuing our goals we often confront challenges that take courage to overcome. We join in friendships with other people and we engage in political activity, both of which require that we treat others generously and justly.

Aristotle thought (and modern-day virtue ethicists have followed him on this) that practical wisdom is an extremely important virtue because it is required to discern the standard of rightness or appropriateness required by any virtue. For Aristotle, practical wisdom requires the other virtues (*Nic. Ethics*: 1144b30) and as soon as someone has practical wisdom, "which is a single state, he has all the virtues as well" (*Nic. Ethics*: 1145a2). He held what has been called the thesis of the "unity of the virtues." The attraction of this thesis is that full virtue seems to require a grasp of the reasons for one's actions (as opposed to merely acting as one has been taught), and the reasons for one's actions bring all the virtues together. For example, the decision about whether to lend money to a friend in need may invoke reasons of justice, compassion, helpfulness, and temperance. To understand what to do (to feel and act appropriately), the practically wise person needs to understand the demands of all the virtues, not just one.

Contemporary virtue ethics has followed Aristotle in thinking that virtues are an essential part of human flourishing or a good human life. Notice that we could agree with Aristotle about this even if we do not agree with him that we have a natural *function* that determines what it is to live a good life. Many Aristotelians today think that the notion of flourishing that includes virtue is not an empirical fact about the natural world, but an ethical ideal (Hursthouse 1999). Contemporary virtue ethics has also been strongly influenced by Aristotle's views about what kind of state a virtue is. There are two interesting features of the Aristotelian conception of a virtue that we will consider in more detail. First, virtues seem to comprise understanding and emotion at the same time. Second, virtues are stable and reliable states that we cultivate in similar ways to how we develop skills. We'll take up the first point in the rest of this section and then turn to the second point in the next section.

On the first point: thus far, we have considered two basic ideas about moral motivation. One is that moral motivation must stem from our desires, because all action is motivated by desire. The other is that moral judgment has a special role in moral motivation. Among those who think moral judgment motivates us to act morally, there are those who think moral judgment is really an expression of sentiments and those who think that moral judgments are judgments about rational requirements. Notice that so far, all the positions we have talked about accept the basic dichotomy between beliefs and desires. Some virtue ethicists have proposed a third alternative, according to which the distinction between beliefs and desires is called into question.

Recall from Chapter 5 the idea that beliefs and desires are distinguished by their different directions of fit. Beliefs aim to fit the world; desires aim to get the world to fit them. Some virtue ethicists take virtues to be states that have both directions of fit simultaneously. For example, a virtue such as generosity is in part a perceptual state that represents facts about the world as considerations in favor of helping others who need help, but it is also a state that moves us to help others when we perceive those considerations. Indeed, in this view, the two directions of fit are intertwined so that if you are not actually moved to help the needy person, then you do not see the reason to help them. To truly perceive that someone else needs help is to be motivated to help. This idea evokes the Socratic thesis that knowledge is sufficient for virtue. In this view, virtue ultimately stems from contemplation of the very idea (or "Platonic form") of the good (Chappell 2014).

There is something intuitively plausible about this picture. If you think about ordinary cases of unethical behavior, it often seems like we can describe the person who acts badly as "not getting it" or "just not seeing what to do." Take the example of Emma from Jane Austen's novel of the same name. Emma is a basically nice, but rather spoiled young woman who is the wealthiest, most privileged person in her little community. At one point she goes on a picnic with a bunch of people, including a flirtatious young man and an "old maid," Miss Bates, an old family friend who is very poor and disadvantaged. At one point the flirtatious young man suggests playing a game in which everyone must say one very clever thing, two moderately clever things or three things "very dull indeed" to amuse Emma (not much of a game, but this was the nineteenth century, after all).

> "Oh! very well," exclaimed Miss Bates, "then I need not be uneasy. 'Three things very dull indeed.' That will just do for me, you know. I shall be sure to say three dull things as soon as ever I open my mouth, shan't I? – (looking round with the most good-humoured dependence on every body's assent) – Do not you all think I shall?"
>
> Emma could not resist.

> "Ah! ma'am, but there may be a difficulty. Pardon me – but you will be limited as to number – only three at once."
>
> (Austen 2000: 242–243)

It becomes very clear that in making this remark Emma has hurt Miss Bates deeply, and Mr. Knightley (the one Emma really wants to impress) reprimands her: "How could you be so insolent in your wit to a woman of her character, age, and situation? – Emma, I had not thought it possible … . It was badly done, indeed!"

Eventually, Emma gets it; she "was forcibly struck. The truth of his representation there was no denying. She felt it at her heart. How could she have been so brutal, so cruel to Miss Bates?" (2000: 245).

The telling thing about this case is that, initially, Emma didn't see what she did as cruel; she saw it as witty. She did not see that the joke she made was truly mean and at Miss Bates's great expense. It isn't that she had bad desires; she did not *want* to make Miss Bates feel mortified; rather, she wanted to be funny and charming. But she misread the situation and thereby ended up acting very badly indeed. Moreover (and I'm interpreting a bit here, but I have read the book many times), had Emma seen the situation rightly, she would not have insulted Miss Bates; it was her perception of the situation and her own role in it, not her desires, that was responsible for her bad behavior.

The virtue ethicist we have been considering would say that Emma lacks virtue because she does not perceive the relevant considerations in the right way, that is, as decisive reasons for acting with kindness toward Miss Bates. But what about the idea that beliefs (or perceptions) and desires (or motivations) must be distinct mental states because each has a different aim? Recall from Chapter 5, if you *believe* that there is an apple in front of you, you will continue to believe it when there is an apple and stop believing it when there isn't, but if you *want* an apple, you will stop wanting it once you've eaten it and continue wanting it when it isn't there. Margaret Little argues that this line of thought does not prove that there couldn't be a mental state (such as a virtue) with two directions of fit. A virtue could be, according to Little,

> a state with two complex properties: it is a believing-attitude directed toward one proposition, and it is a desiring-attitude directed toward another. There is nothing formally odd in saying that a belief(p) can also be the desire(q), just as there is nothing formally odd in noting that the mathematical operation "add(2)" is also the operation "subtract(-2)."
>
> (Little 1997: 64)

These states with double directions of fit are sometimes called "besires" because they seem to have some features of beliefs and some features of desires (Altham 1986).

At first glance, there does seem to be something wrong with the idea of a "besire." We can certainly imagine two people who believe the same facts about a situation, one of whom is motivated to do the right thing and the other of whom is not. How could the fact that the second person doesn't have the right motives take away from what she knows about her situation? The Aristotelian thinks this is the wrong way of looking at it. The knowledge that is inherent in virtue is not that kind of knowledge; it's not like knowledge of a fact. Rather, it is more like perceptual knowledge; it is the discernment of the morally salient considerations in a situation, where "'taking as salient' is akin to having a kind of experience" (Little 1997: 73; see also Nussbaum 1985). The person who isn't motivated to help the needy person, according to this account of the virtues, has a cognitive defect, but it is a different kind of defect from the possession of a false belief; she doesn't see the situation correctly. If you've ever been in a situation where you've watched someone behave really badly – insulting a friend, telling a racist joke, humiliating a colleague – this should sound familiar. Sometimes you find yourself unable to believe that the person you're watching could have such a poor grasp of the situation. You may not think the insulter or humiliator wants to be mean, but that he or she is (cognitively) clueless. Indeed, a modern movie version of Jane Austen's *Emma* is called *Clueless*. Of course, the defender of the view that virtue is like perception thinks that *every* time someone acts badly it is because they are clueless. This is a strong claim, and you might think of some counter-examples, but I hope the examples we've considered allow you to think of what this view might have going for it.

One thing we might wonder about this position on the virtues, given our previous discussions, is whether it provides any traction in the debate between sentimentalism and rationalism about moral judgment. This is not so clear. Recall that this is a debate about whether moral judgments are expressions of sentiments or rational judgments based on principles. One thing that was attractive about rationalism was that it put our rational capacities in charge in a special way: according to the Kantian picture, the rational capacities that are in the driver's seat are responding to principles that justify our actions. The worry about sentimentalism was that if the sentiments are in charge, then we're really just doing what we feel like doing and true justification goes out the window. Does the Aristotelian virtue ethicist give us a good alternative? Aristotelians do not think that there are universal rational principles that explain why actions are right or wrong. However, virtue ethics does make sense of the idea that moral reasons are there for all of us to perceive. According to virtue ethics, we are not just doing what we feel like doing when we act morally; we are grasping the ethically salient features of a situation and responding appropriately to them by using our rational capacities.[1]

In this section we have seen that some virtue ethicists take virtue to be a unique kind of motivation that involves both thinking and feeling. If there are such states, they would help to explain the distinctiveness of moral motivation. The existence of such states is controversial, however, and not all virtue ethicists accept them. It is open to a virtue ethicist to think that virtues are complex traits, constituted by beliefs, desires, and emotions that are only contingently related. This picture of what a virtue consists in would mesh nicely with the conception of a trait in personality psychology according to which

> [p]ersonality traits are probabilistic descriptions of relatively stable patterns of emotion, motivation, cognition, and behavior, in response to classes of stimuli that have been present in human cultures over evolutionary time.
> (DeYoung 2015: 3)

In the next section we'll turn to a different aspect of the virtues – their alleged stability and consistency across different situations – that many virtue ethicists accept, whatever they think about beliefs and desires.

Should We Aspire to Virtue? The Empirical Challenge

As I said at the beginning of the chapter, one of the attractive things about virtue ethics is that character evaluations are so natural to us. Nevertheless, there is some research that indicates that character attributions are not very reliable, and a serious attack on virtue ethics has been launched on the basis of this research. In fact, our tendency to focus on internal causes of people's actions rather than powerful situational causes has been called a kind of error: the fundamental attribution error (Ross 1977). It seems that when we try to explain why people do things, we tend to focus on the personality, character traits, and motives of the actors, rather than on external, situational factors. There is experimental evidence that we do this even in cases in which it should be obvious that the situational explanation is the right one. For example, in one experiment participants were asked to assess people's true attitudes toward Fidel Castro based on an essay these people had allegedly written about Castro. It's fair enough that if I praise Castro in an essay, you'll think I'm pro-Castro. But in this case, participants were explicitly told that the writers had no choice about their essay topic: they were assigned a position, pro or con, and required to argue that point of view no matter what they actually thought. It is therefore surprising that participants were more inclined to say that someone who wrote a pro-Castro essay was truly pro-Castro, and more inclined to say that someone who wrote an anti-Castro essay was truly anti-Castro (Jones & Harris 1967). This is just one small example, but it turns out that in all sorts of circumstances, we look for explanations in the character of the person and we tend to ignore the external causes of their actions.

Looking to character to explain people's actions would be fine if character were frequently the cause of behavior. But there is reason to be skeptical. Philosophers John Doris (1998; 2002) and Gilbert Harman (1999) have launched a serious attack on the idea that the philosophical picture assumed by virtue ethics is widely applicable to human beings. To understand this attack, we need to talk about another aspect of the philosophical characterization of virtue, which is that virtues are supposed to be stable and consistently reliable states of character.

According to the virtue tradition in philosophy, virtues are states of character that are consistent or reliable in the sense that, once a person has developed virtue, she will tend to act well even in difficult circumstances. When the going gets tough, the truly virtuous continue to act in accordance with virtue. This is certainly part of Aristotle's conception of a virtue. He thought that the actions of the virtuous "proceed from a firm and unchangeable character" (*Nic. Ethics*: 1105a) and that a wise person "will bear all kinds of fortune in a seemly way, and will always act in the noblest manner that the circumstances allow" (*Nic. Ethics*: 1101a). It also seems to me that this is part of our ordinary understanding of virtue. When we say that someone is an honest person, we imply that they are consistently honest, even when telling the truth is hard. A person who was honest with her colleagues at work but lied to her friends probably wouldn't earn the title of honest person. The perfectly virtuous person who perceives the situation correctly will be motivated to act rightly in all circumstances. As Doris puts it, "virtues are supposed to be *robust* traits; if a person has a robust trait, they can be confidently expected to display trait-relevant behavior across a wide variety of trait-relevant situations, even where some or all of these situations are not optimally conducive to such behavior" (Doris 2002: 18).

To put Doris's argument against virtue ethics in the simplest form, the basic idea is that virtue ethics assumes that there are robust, cross-situationally reliable traits (that is, it assumes there are virtues). Research in psychology gives us very good evidence that traits like this are not widely instantiated. Therefore, there is something wrong with virtue ethics. To put Doris's argument in a slightly less simple (and, therefore, more charitable) way, it seems to me it goes like this:

1. A good normative theory must give us something to aim at that will help us behave well.
2. Virtue ethics recommends that we aim at reforming our character in order to behave well.
3. This recommendation presupposes that we (most of us) can develop certain robust, cross-situationally reliable character traits to a reasonable degree (namely, the virtues that will motivate good behavior).[2]
4. Research in psychology provides evidence that such robust traits of character that reliably produce behavior across various contexts are only possessed by the exceptional person; hardly any people have (even a reasonable degree of) such traits.
5. Aiming at developing the virtues is, therefore, not a good way to ensure that we will behave well.
6. Therefore, virtue ethics is not a good normative theory.

To evaluate this argument, we need to see what the evidence is for premise 4. Doris draws on hundreds of studies to make his case. We obviously can't survey all the relevant evidence here. Instead, I'll discuss a few examples to give you the flavor of the research.

Let's start with an experiment that will be familiar to most people, the Milgram experiment (Milgram 1963). In this experiment the subjects were told that they were in a study on learning and that they would be the "teachers" in this study. Their "student" (who was really a confederate of the experimenter, Stanley Milgram) was hooked up to a shock generator, and the subjects of the experiment (the "teachers") were asked to deliver shocks when the student made mistakes. The shock generator had labels for the increasing amounts of electricity from "slight shock" (15 volts) to "danger: severe shock" (375 volts) and then to "XXX" (450 volts, the highest number on the scale). The "student" would pretend to protest the pain (the confederate wasn't really being shocked, of course), sometimes even screaming to be let out of the experiment. When the "teachers" resisted administering stronger shocks, the experimenter would calmly say "please continue" and, ultimately, "You have no other choice, you must go on."

As is well known now, what people did in this experiment was, well, shocking. Sixty-five percent of the subjects delivered the maximum shock.[3] Of the 40 subjects, none stopped administering shocks before the level of 300 volts or the end of the "intense shock" range, the point at which the "student" "kicks on the wall and no longer provides answers to the teacher's multiple-choice questions" (Milgram 1963: 375).

What is most interesting about this, for our purposes, is that it pushes us to question the validity of our own attributions of character. When I first heard about this experimental set-up, I was extremely surprised that the subjects in the experiment would do what they did and I felt quite sure that if it were me I would not have done it. Indeed, if you ask people what they would do, as Milgram did, most say they would stop early in the process and virtually nobody thinks anyone would go all the way to the 450 volt limit. As I thought about my own predictions, I had to admit that I'm not fundamentally different from the people who were in that experiment and that all of those people would likely have thought the same thing about themselves. I would bet that most people reading this (or hearing about the Milgram experiments for the first time) thought: I wouldn't do it! I would protest against the white coat! But only a few people really did resist, and we can't all be the exception. This should give us some pause about how well we are able to discern what our motivational tendencies are.

The influence of an authority figure can make people do terrible things, it seems. If we assume that most people are not entirely lacking in the virtue of kindness, this seems surprising. Another surprising finding, again if we assume people have even a modicum of the virtue of helpfulness, is the bystander effect. The bystander effect is the influence of the presence of other people (passive bystanders) on the likelihood of helping in a situation that calls for help. To investigate the bystander effect, researchers put people in a situation in which someone appears to need help: one of the researchers falls off a book shelf, has a seizure or an asthma attack, or looks to be in some other way in need of assistance. The study participants are either in a situation with other people who are passive bystanders or a situation with no one else or just one other person. Then the researchers look to see who helps and how long it takes them. In an early review of this research, Latané and Nida (1981) found that it makes a real difference how many people are there when you witness someone in need: overall, three quarters of the participants helped when they faced a critical incident alone, only a little more than half helped when other people were present. A more recent meta-analysis confirms the basic picture, providing "clear support for the assumption that passive bystanders in critical situations reduce helping" (Fischer et al. 2011).

Much of the classic research on the bystander effect took place in the late 1960s and early 1970s. Perhaps in our social media age, we are not surprised by the negative effects of crowds. And there has been research on the bystander effect in social media. In one such study, from the early days of social media, Patrick Markey (2000) investigated 400 different chat groups that ranged in size from 2 to 19 people. Markey sought help in each group, simply by asking the question "Can anyone tell me how to look at someone's profile?" The larger the group, the longer it took anyone to respond to Markey's question. (Interestingly, Markey found that it made no difference whether the person asking for help had a male sounding name or a female sounding name, but it did make a big difference if the question mentioned someone else's name. People were more likely to help, regardless of the size of the chat room, if they were called on by name).

These are just a few of the many studies that have demonstrated the strong influence of the situation on our behavior. Doris argues that this evidence weighs heavily against the idea that very many of us have anything like a reasonable degree of virtuous traits (like compassion and helpfulness) that reliably influence our behavior independently of the situational factors. Notice that given the way that I have reconstructed Doris' argument, Doris does not need to say (nor does he) that virtues are not possible. Rather, the claim he needs for his argument to work is that the development of such traits is unlikely enough that it doesn't make sense for an ethical theory to be organized around the cultivation of virtue. He says, "The situationist does not deny that people have personality traits; she instead denies that people *typically* have highly general personality traits that effect behavior manifesting a high degree of cross-situational consistency" (Doris 2002: 39, italics added). As far as the evidence goes, Doris accepts that there may be "local" traits (e.g., honesty with the boss when I'm not in any trouble), and he admits that there may be atypical cases of people who have "global" traits that really are firm and unchanging, even when the person is pressured by a scientist or surrounded by passive bystanders. The point is that even if there could be a few people in the world whose personality goes a long way toward making them behave well, situational factors are such a strong influence on the rest of us that working on our character doesn't seem a very good strategy for moral improvement.

The conclusion of Doris' argument, as I've reconstructed it, puts virtue ethics into some disrepute. You might now be wondering about the repute of the psychological research Doris relies on. As I mentioned in the introductory chapter, a good deal of research in social psychology has been challenged by the replication crisis or "repligate," and some key results of past research are under a cloud. Indeed, some of the studies of situational influence I discussed in the previous edition of this book have been shown not to hold up to replication.[4]

The issue here is complex and much has been written about it, especially by Doris himself (2015; 2022). The short version of how things stand is that even if not all of the studies purporting to show the influence of situation really demonstrate what they claimed, it doesn't really matter for the influence of character, because everyone agrees that character is just one (not huge) influence on our behavior. Moreover, situation is also just one influence on our behavior. The conclusion to draw, it seems to me, is that if we want to use an ethical theory to suggest ways for us to behave better, we shouldn't rely on an outmoded conception of virtue that takes virtue to be like an impenetrable shield against bad behavior, but that doesn't mean that there's nothing about character that will help us.

The point I just made – that character is just one (not huge) influence on behavior – is important and needs some explanation. The point is about the "effect size" in the research on the relative influence of traits and situational factors. An effect size is a number that expresses the strength or magnitude of the relationship between two variables. There are a few things to keep in mind about effect sizes. First, notice that even though the word "effect" is used, effect sizes do not by themselves report causal relationships; much of the research here is correlational research. Second, effect sizes are between 1 (the strongest possible relationship: perfect correlation between the variables) and zero (no relationship between the variables). What everyone seems to agree about is that it's seldom reasonable to hope for an effect size greater than about .3 for the relationship between personality or situational factors (on the one hand) and particular behaviors (on the other hand) (Doris 2022). The reason for this is just that human behavior is incredibly complicated and many factors contribute to explaining it (Ahadi & Diener 1989). It's a consensus in personality and social psychology right now that both personality traits and situation influence human behavior. Third, effect sizes are standardized so that meaningful comparisons can be made across different variables. (Say, for instance, the number of passive bystanders and the time it takes for a person to help.) What this means is that in order to know what an effect size really means, you have to know something about the scales that are used in the experiment.[5]

Focusing on character again, is an effect size of .3 big or small? It's going to depend on what you're measuring and for what purpose. In personality and social psychology, there are conventions for this according to which an effect size of .1 is small, an effect size of .3 is medium, and an effect size of .5 is large (Cohen 1988). The more important question for our purposes, is this: is an effect size of .3 for the effects of character traits or virtues on behavior *big enough* for virtue ethics? Doris (2022) argues that the important point is that an effect size of .3 is smaller than what we would expect for virtue. "The way I'd put it now," he says, "is roughly this: in many cases, situational variables matter more, and personality variables less, than one might expect (2022: 217).

What does this mean for virtue ethics? It depends on what virtue ethics assumes and what it hopes to accomplish. If virtue ethics assumes that most people can develop character traits that reliably produce good behavior and dependably prevent bad behavior, regardless of the situation, virtue ethics is on shaky ground. But some virtue ethicists have found firmer ground to stand on.

Defending Virtue

Doris' critique of virtue ethics ruffled a lot of feathers and has received a good deal of attention. Some defenders of virtue ethics have responded to the situationist challenge by saying that it misses the target of the traditional conception of virtue, so virtue ethics can go on just as before. Others have responded by modifying the conception of virtue (or advocating a different traditional conception; e.g., one from Hume rather than Aristotle). In this section we'll consider these different ways of responding.

Some philosophers have argued that Doris' critique isn't fair to virtue ethics because it does not give enough attention to the role of practical wisdom in virtue and that, once this role is appreciated, the situationist critique no longer applies. Rachana Kamtekar, for example, argues that the empirical attack on virtue ethics misses the target, because the experiments that are supposed to show that character traits do not explain our behavior rely on a faulty conception of a character trait "as an isolable and nonrational disposition to manifest a given stereotypical behavior that differs from the behavior of others and is fairly situation insensitive" (Kamtekar 2004: 477). But according to virtue ethics in the tradition of Aristotle, Kamtekar says, virtues are "dispositions to respond appropriately – in judgment, feeling, and action – to one's situation. Such responses require the active involvement of the agent's powers of reasoning" (2004: 477). In other words, as we discussed in the previous section, virtues require practical wisdom if they are going to help people behave well. Since there is no reason to expect the average participant in a psychological study to be wise, such studies do not show that virtuous people do not behave well even when they confronting an authority, or surrounded by passive bystanders.

It's not clear how much this response helps to vindicate virtue ethics, if the argument I attributed to Doris above is correct. That argument takes the problem to be that cross-situationally reliable virtues are not widely instantiated in the population, that is, not widely instantiated enough to make cultivating virtue a good strategy for moral improvement. Practical wisdom seems to be quite rare and difficult to cultivate. If the full possession of any virtue requires practical wisdom, then it will be the rare person who really does possess virtue (a fact which Kamtekar admits). If Kamtekar is correct about the importance of wisdom, and if normative theories are to be judged on whether they provide good strategies for moral improvement, then whether virtue ethics is a good normative theory will depend on whether we can make strides in improving our practical reasoning and whether wise reflection can help us make better choices.[6] Whether it makes sense to try to cultivate wisdom and the other virtues along with it depends in part on how practical wisdom is conceived.

We will come back to practical wisdom in a moment, but first let's consider a different approach. Instead of defending the traditional, Aristotelian conception of virtue, you might start with what we know about traits from psychology and build a conception of virtue out of that. As character skeptics like Doris admit, traits do influence behavior – studies of the influence of personality do show character traits influence behavior, even though they are not the only influence. And as I mentioned above, psychologists agree that situation *and* personality both have a role to play in determining behavior. When it comes to the personality side of things, psychologists tend to think there is very good evidence for the "Big Five" personality traits: Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion, (the list forms CANOE, if you want to remember it) (John, Robins, & Pervin 2008). These traits are not moral virtues, but the fact that psychologists think there are such traits and that they influence behavior across situations does open some opportunities for defending character traits that are virtues.[7]

How is the existence of these stable personality traits compatible with all the evidence that situational factors have tremendous influence? One answer is that although situational factors have an influence in particular instances, we see the influence of traits when we look at patterns of behavior over time.[8] The idea here is that traits are patterns of psychological dispositions, or as William Fleeson and colleagues put it, character traits are "density distributions" (2001). Basically, according to the density distribution approach, while it is true that an individual person's behavior varies from situation to situation, the various behaviors of one person form a distribution that has a central tendency, and we can think of a person's character as defined by these central tendencies within distributions. Fleeson and his colleagues have found that people really do differ from each other in terms of what we might call their "average behavior" and that these differences are very stable (Fleeson 2001). Furthermore, though most of the research has been done on the Big Five traits, these psychologists argue that their model is a good model for understanding virtue (Jayawickreme, Meindl, Helzer, Furr, & Fleeson 2014). At the very least, this gives us a way to understand what a trait is that is compatible with the research on situationism, and this opens up a new line of response for the virtue ethicist.

To see how this would work in more detail, consider Figure 9.1. The figure shows density distributions of compassionate behavior for fifty hypothetical people.[9] Nobody is always compassionate, and how compassionate a given person is varies from situation to situation, but people vary in different ways. A person whose shaded area is weighted toward the right is more compassionate more often; a person whose shaded area is weighted to the left is often not compassionate and rarely very compassionate. If we think of virtues as density distributions, we would say that the right-weighted people are more

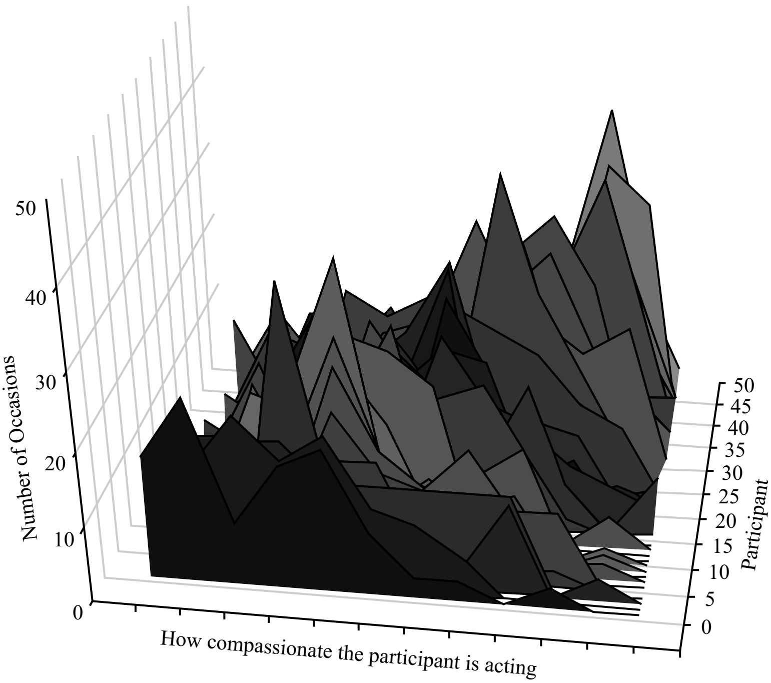


Figure 9.1 Distributions of Compassionate Behavior.

compassionate and the left-weighted people are more heartless. There are also people in between who are not quite compassionate or heartless.

This understanding of traits, as behavioral averages or density distributions, is compatible with thinking that situations can make a big difference.[10] If we think of virtues as traits of this kind, then virtue ethics can admit that virtues are frail or wobbly as long as it still makes sense to praise people for their "central tendencies" and aiming to improve your central tendency still constitutes a reasonable ethical goal.[11] The philosopher Christian Miller (2013) has developed a theory of cross-situationally consistent character traits that fits well with this psychological research. Miller argues that what we human beings have are "mixed traits"; we tend to act well in some circumstances and to act badly in others, just as the empirical studies reveal. Virtue ethics can be built on mixed traits as long as we can set about to improve the balance of good to bad.

What virtue ethics will have to give up on, according to Maria Merritt (2000), is the Aristotelian idea that virtue implies very strong "motivational self-sufficiency," that is, the power to motivate us all by themselves without any help in whatever tempting circumstances we find ourselves. "What situationist psychology makes problematic," according to Merritt, "is not as such the recommendation to have the virtues, but the normative ideal of the virtues as qualities that must be possessed in a strongly self-sufficient form" (2000: 375).

In short, we can move our character from good to bad, according to Merritt, but only with the help of supportive institutions and communities.

Merritt's proposal is to think of virtues as David Hume did: traits that are relatively stable and reliable at producing good actions, but not necessarily because of an internal psychological force. Following Hume, virtue ethics could conceive of virtues as whatever habits and dispositions get us to behave better, where these habits and dispositions might require specific situational reinforcement. On this view, we should appreciate the ways in which virtues are developed and maintained within social structures and acknowledge the importance of these structures in upholding virtue. Thinking about virtue in this way, we could say that what went wrong in the Milgram experiments is that people were placed in an unusual social environment in which the normal support for compassion was absent and the influence of the authority figure was highly salient. A Humean virtue ethicist should be particularly interested in making sure that our social environments sustain and facilitate our acting on our good traits.

Another approach to thinking about how we can move ourselves toward the more virtuous end of the bell curve is the skill model of virtue. The skill model of virtue has its roots in Aristotle and also in ancient Chinese philosophy (Mengzi, in Van Norden 2008), but recent defenders draw on psychological evidence about skill acquisition and expertise to elaborate and defend it. The basic idea is that virtues are analogous to the skills necessary for playing tennis, chess, or cello. Aristotle thought that we have natural capacities for virtues that we must "activate" and train through habituation. "For we learn a craft by producing the same product we must produce when we have learned it; we become builders, for instance, by building, and we become harpists by playing the harp. Similarly, then, we become just by doing just actions, temperate by doing temperate actions, brave by doing brave actions" (*Nic. Ethics*: 1103a31–1103b2). Given Aristotle's conception of virtue, what we are supposed to learn, ultimately, from habituation is to think, feel, and act appropriately from a settled disposition.

The idea that you can become virtuous yourself by doing what the virtuous person does has some common-sense appeal. I know that I have sometimes looked to people I admire for their generosity or their conscientiousness and emulated what they do, and we do tend to think that if you talk the talk for long enough, you'll end up walking the walk. But the skill model also raises some questions. First, if emulation is a key part of training, how do we identify the virtuous person? We need to know whom to copy if copying is going to make us better. Second, is emulation sufficient? It certainly isn't in tennis: you can't become a great tennis player by trying to copy Serena Williams and hoping for the best.

The first question about exemplars is really two questions: how do we characterize or define the fully virtuous agent philosophically? And how do we identify a virtuous agent in practice? The first question takes us deep into virtue ethics and will be answered differently by different philosophers. For Aristotelians, however, the answer will make reference to human capacities and flourishing. The answer to the second question is likely going to be that we engage in a process where we identify good people, become better ourselves, which allows us to refine our sense of what a good person is, which in turn allows us to emulate better people, and so on (see Zagzebski 2004).

The second question about the specific skills involved in cultivating virtue has received a lot of attention from empirically minded philosophers, and one of the main skills that has been identified is self-regulation. The skills of self-regulation enable the effective pursuit of goals; they allow people to change how they think, feel, and behave in ways that better conform to desirable standards. For example, a person who is good at self-regulation will be better able to work despite distractions, control their impulses for unhealthy foods or cigarettes, and calm themselves in a stressful situation (Baumeister 2022). Virtues require conforming to moral standards, or to the standards of a virtuous exemplar, so, as Lorraine Besser puts it, we can see a virtuous person as "a self-regulated agent, who successfully regulates herself by virtue-related goals, using whatever strategies prove effective in the context" (Besser 2022: 167; 2017).

Besser (2017) highlights two specific skills that are part of the self-regulation suite: the capacity to form implementation intentions and understanding of the hierarchy of goals. Implementation intentions are specific action-guiding plans that further one's goals. If your goal is to be more conscientious about studying, you are better off having a specific plan for how to do this – work in the library and shut down your browser every Tuesday and Wednesday evening, for example – than to have the vague goal "study more!" Understanding the way goals are organized is also important. You need to know that your goal is to learn more or to do better in your classes in order to come up with a good plan. If you think your ultimate goal is just to study more, you won't be sensitive to the information that studying more isn't working (perhaps you need to study differently).

Matt Stichter, another advocate of the skill model, agrees that the skills of self-regulation are crucial for the development of virtue (2007). In addition to the skills we've already discussed, Stichter (2020) highlights the importance of *emotional* self-regulation and the skill of emotion differentiation, which is the ability to translate a vague mood into a specific emotional state. He argues that people who can be specific about what they are feeling are better able to manage those feelings and respond to them appropriately, which in turn makes them better able to learn from their emotions in the process of cultivating a virtue.

To get a "feel" for the point about emotion differentiation, consider one of my fellow graduate students, Jack (not his real name), whom I observed as he played beach volleyball with the other sporty grad students. When Jack would miss an important shot, he would jump up and down in the sand screaming "I suck! I suck! I suck!" It was quite dramatic, but spectacularly unhelpful. Jack would become distracted by his negative mood, drink more beer, and play worse and worse, resulting in more and more yelling. This is not a good state to be in for learning from one's mistakes! A generalized self-hating mood is demoralizing and it closes a person off from constructive criticism or coaching.

Jack would have been better off if he could have seen that he was feeling angry at himself for missing a shot that he thought he should have made.

What does this have to do with virtue? Developing the skills of virtue will require learning from our mistakes. If you want to be more conscientious about studying, you will sometimes fail and you need to improve your plans based on those failures. You won't do that if you're paralyzed by shame or self-loathing. Furthermore, many virtues require the regulation of emotions as a part of the virtue. As the well-known (and previously quoted) passage from Aristotle puts it, "the person who is angry at the right things and toward the right people, and also in the right way, at the right time, and for the right length of time, is praised" (*Nic. Ethics*: 1125b31–33).

The skill model has some advantages. It makes sense of why there aren't that many perfectly virtuous people: expertise is difficult to cultivate and there aren't very many experts in other domains either. It also makes sense of why it still makes sense to think about virtues as an ethical goal: we can get better, even if we can't be perfect, and that seems ethically important. Moreover, the skill model combines nicely with the findings from social psychology that trouble the character skeptics. Knowing the tendencies we have to be influenced by the wrong things – authority figures, the presence of bystanders, selfish bias, and so on – can help us in our training, just as knowing that you tend not to follow through in your tennis stroke can help you play better. The skill model also combines nicely with Merritt's insights about the importance of social support (see also Ciurria 2014). After all, social support is crucial to developing other skills – we need good teachers and coaches, good environments in which to practice, other musicians or athletes to play with and reinforce our good habits.

At the beginning of this section, I said that some philosophers argue that the situationist research in psychology misunderstands the nature of virtue because it ignores the role of practical wisdom in virtue. It was left unclear whether developing wisdom is a useful ethical goal. We can now see what the problem is more precisely: if practical wisdom requires strong motivational self-sufficiency, then developing it would be just as problematic as developing any other kind of self-sufficient virtue. But does practical wisdom necessarily include a high degree of motivational self-sufficiency? Or can we think of wisdom as a socially supported set of skills, similar to other virtues?

The wise person, according to Aristotle, "is the one whose aim expresses rational calculation in pursuit of the best good for a human being that is achievable in action" (*Nic. Ethics*: 1141b10–15). It is true that the traditional conception of wisdom takes it to be rather autonomous or self-sufficient. We talked earlier about philosophers who liken virtue to a kind of perception that is all you need for moral motivation, and this view may also apply to wisdom. The philosopher John McDowell, for instance, takes practical wisdom to be analogous to a kind of sensitive perception of what the situation requires, which results in the fully virtuous person seeing a course of action as "the thing to do" in a way that is sufficient for her to do it. According to McDowell, "nothing over and above the unclouded deliverances of the sensitivity is needed to explain the actions that manifest virtue" (1979: 334). This does seem to be a picture of wisdom that ignores what Merritt calls the social contribution to virtue. It is also a picture of wisdom according to which to truly achieve wisdom requires quite a bit more control over how we see things than we might actually have.

Do we have to see practical wisdom this way? Not necessarily. We could agree with Aristotle that wisdom is the ability and the will to deliberate well about the ends of human life and action, but allow our particular view about what skills are required by this good deliberation to be informed by a realistic understanding of our psychology. For example, Jason Swartwood (2013) draws on the Recognition-Primed Decision (RPD) model from the psychology of expert decision-making to identify the skills that constitute wisdom. According to this model, expertise is an understanding of how to conduct oneself in a particular domain. Wisdom, according to Swartwood, is the understanding of how to conduct oneself all things considered and it uses the same kinds of skills. What are those skills?

> The first is *intuitive ability*: experts can identify what ought to be done quickly, effortlessly, and without conscious deliberation. The second component ability is *deliberative ability*: experts use effortful, consciously accessible processes to search for and evaluate choices when an intuitive identification is lacking or inadequate … A third component ability implied by the RPD model is *meta-cognitive ability*: an expert is able to decide when and how to rely on intuition and deliberation.
>
> (Swartwood 2013: 518)

Because the domain of wisdom is so complex – it encompasses everything, after all! – two more skills are needed: self-regulative ability and self-cultivation (both of which we have already touched on in our discussion of the skill model above).

To see how these skills might work in practice, think back to the trolley problems from Chapter 8. Imagine a person with developed skills of wisdom who is faced with the choice between pushing a large man in front of a trolley, thus saving five other innocent lives, or letting the trolley continue on its way, killing five but avoiding the large man. Using their intuitive skills, the wise person will understand that pushing people into trolleys is a no-go, morally speaking. However, their meta-cognitive ability will allow them to recognize that this is an unusual situation in which conscious deliberation might be called for. The wise person will be able to regulate their emotional responses to the situation, so that if they decide that they ought to push the large man they will understand and respond appropriately to their sadness and regret. This does seem like a reasonable approximation of the process that a wise person would use to decide.

My own thoughts about what I have called "reflective wisdom" (the kind of wisdom needed to live our own lives well) has also been informed by research on our psychological limitations. In a book with the subtitle *living wisely with our limits*, I point out that we have poor self-knowledge, we are bad at introspecting our own motivations for doing things, and we have a tendency to become obsessed with trivial things and to ignore what's important until it is threatened (Tiberius 2008). All of these things make us bad at practical reasoning because they distort the inputs to the process. I argue that, to improve our deliberation, what we need to do is to cultivate some background dispositions and habits that will make it less likely that we'll reason badly. I call this set of dispositions "reflective wisdom," which includes perspective, self-awareness, and the development of stable, emotionally appropriate values. The dispositions and habits of mind that constitute reflective wisdom include paying attention to the situation and relying on other people (for instance as a source of self-knowledge) to help compensate for your limitations.

We have arrived at a conception of virtue that needn't be too embarrassed by the small to medium-sized relationship between character and action. As in tennis, if I can get some gains by practicing, even though I will never play like Serena, I should do that (assuming that I want to be a better tennis player). Developing the skills of virtue won't make us morally perfect, but it seems reasonable to think it will make us morally better. Doris (2022: 249) cautions against continuing to include virtue in our ethical toolkit. He thinks we would be better off focusing on rules and principles, given our tendency to make too much out of character (the fundamental attribution error), which gives rise to unrealistic expectations and harsh appraisals of others. I think Doris is surely correct that virtue isn't the only ethical ideal we should consider in our moral practice. Focusing on rules and principles makes sense, too, especially given our discussion in Chapter 2 of the ways in which we evolved to have capacities to internalize rules and reason about them together. But should we avoid talking about virtue altogether? Perhaps these tendencies to inappropriate and harsh generalizations about people can be mitigated by wisdom and compassion. Can they? Well, it seems like the answer to that question is: somewhat.

Taking Stock

It's worth asking what is at stake in the debate about virtue. "Virtue talk" is ubiquitous. Voters obsess about the character of political candidates; grade-school teachers plan how to inculcate the virtues of honesty and kindness in their students; all of us judge the people we meet according to how nice, mean, dependable and so forth they are; parents advise us to date and marry nice people; we make New Year's resolutions to "be kinder" or more generous; and so on. If we cannot find a way of thinking about virtue, wisdom, and practical reasoning that respects our psychological reality, none of this makes sense.[12] Many of us have judged people (including ourselves) unfairly: we say that someone is mean, but really he was just having a bad day; we beat ourselves up for being too lazy to stick to some new exercise regimen, but really it's the fault of our sedentary culture. One of the things we should learn from the situationist critique of virtue ethics is that it doesn't make sense to ignore situational factors and pretend that people have perfect control over how they behave, or that we can just will ourselves to be better people. An integrated conception of character that takes in the insight from situationism would not recommend that we judge people harshly regardless of situation or that we try to change ourselves by sheer force of will. Rather, it would recommend directing our attention to the ways in which situations affect us and others. We should take this into account when we judge others, and we should use our reasoning to figure out the best way to overcome the pressures and temptations that we have good reason to want to overcome. Of course, even this kind of reasoning might be ineffective and even this kind of virtue might elude us, but in my view we would need a much stronger argument than what we have now to make it reasonable to give up on these hopes altogether.

If we can find a psychologically defensible theory of virtue, what does this tell us about moral motivation? If virtues are taken to be truly special states that are neither judgments nor desires, then moral motivation will be something quite unique, according to virtue ethics. If, on the other hand, moral action, according to virtue ethics, is action motivated by relatively stable psychological patterns that include emotions, desires, judgments, and the commitment to doing what it takes to be a good person (including being careful about your situation), then the virtue ethical picture of moral motivation is not really incompatible with the other theories we have seen so far. Rather, virtue ethics just picks out which specific desires, sentiments, and judgments are essential to moral motivation, that is, the ones that make up the virtues and are appropriately related to our flourishing.

Summary

- Virtue ethicists claim that moral motivation is a distinctive kind of motivation. We are motivated to act morally by virtuous character traits.
- Virtues are identified by their relationship to human flourishing or happiness. Aristotle takes virtuous activity to be the essence of a flourishing life for a person.
- It has been argued that virtues are sensitivities to moral considerations that have both directions of fit: they are perceptions of reasons to do things that also motivate us to act. This is a controversial view about virtue.
- Less controversial is the claim that virtues are stable and consistent traits of character. This claim has been challenged by empirical research on the effect of situational factors on our behavior (the situationist challenge).
- The tremendous and surprising influence of situation has been taken by critics of virtue ethics to undermine the idea that cross-situationally reliable, motivationally self-sufficient traits are possessed by anyone but the rare and exceptional person.
- It would be bad for virtue ethics if virtues are so rare that aiming to cultivate virtues is not a good avenue to moral improvement.
- One response to the situationist challenge has been to argue that it misses the target because it ignores the role of practical wisdom in the traditional conception of virtue.

- Another response has been to construe virtues in a different way so that they are not motivationally self-sufficient, but to argue that they still constitute a good ethical goal.

Study Questions

1. Thinking about the people you know, how stable do you observe character to be? Think about someone you think is a really good person: Is this person good in every domain? Do they have all the virtues? Do they exhibit the virtues all the time? Now think about someone who's kind of normal (in terms of moral goodness). What would you say is the relevant difference between these two people?
2. Sometimes in moral psychology, progress is made by identifying an empirical assumption that some moral theory makes and then showing how this assumption is undermined or supported by the empirical evidence. Has this kind of progress been made with respect to virtue ethics?
3. Politicians often get caught cheating on their wives (or, less often, their husbands), having affairs with young interns, and in various other ways flouting their own professed sexual morality. The public tends to judge them harshly, and they often end up getting booted out of office. Should the research on the power of the situation make us (the public) more lenient and less judgmental?
4. In the business world, there is an acronym for the kind of goals we ought to set if we want to be successful: Specific, Measurable, Attainable, Relevant, Time-Bound (SMART). Think of an example of a person who aims to cultivate a virtue. Is this list of features helpful. Is there anything you would add or subtract?

Notes

[1] How virtue ethics should defend the metaethics that this picture requires – a worldview according to which it makes sense to say that ethical appreciation is akin to perception – is an interesting question that we won't go into here.

[2] Doris' argument against virtue ethics assumes only that virtues can be developed by most people "to a reasonable degree," not that we must all be able to achieve perfection.

[3] This was in the original version of the experiment. Milgram ran many versions in the early 1960s and compliance varied from very little to almost total. For discussion of the follow-up experiments, see Milgram (1974).

[4] The dime in the phone booth study (Isen & Levin 1972), in particular, has been criticized. See the discussion in the Appendix of Christian Miller (2003) for an early critique of this research.

[5] If you want to understand this better, you can start with Karen Grace-Martin's explanation of effect sizes here: https://www.theanalysisfactor.com/two-types-effect-size-statistic/. The Analysis Factor is a great website for basic explanations of statistical concepts.

[6] Doris (2009) has come back fighting against this response to his critique, by arguing that we do not actually fare very well when we let our rational capacities take charge.

[7] There is also evidence that we can change our character traits, at least a little bit. See Edmonds et al. (2008).

[8] Nancy Snow (2010) takes a different empirical approach, grounding the virtues on the Cognitive-Affective Personality System (CAPS) model in psychology.

[9] Thanks to William Fleeson and Eranda Jayawickreme for permission to use their figure. The point of the figure is to present a model for the virtue of compassion; it represents hypothetical patterns of behavior based on patterns tested for other traits.

[10] It is also compatible with the definition of traits from personality psychology as probabilistic descriptions of psychological patterns, mentioned earlier in the chapter.

[11] Doris suggests that these psychologists and the philosophical character skeptics may be talking past each other (2022: 229). The psychologists are pointing to between-person consistency, while the philosophers are looking for within-person consistency. In the final part of this chapter, we consider the idea that between-person consistency is enough provided that a person can move her character profile toward a *more* virtuous distribution.

[12] Not everyone agrees. Mark Alfano (2013) argues that we should keep virtue in our ethical toolkit, even if attributions of virtue don't make psychological sense. He argues for an error theory of character, according to which we should carry on as if there were virtues (even if there aren't), because the language of virtue is often a self-fulfilling prophecy.

Further Readings

Besser, L. 2017. Virtue of self-regulation. *Ethical Theory and Moral Practice 20*(3): 505–517.

Doris, J. M. 1998. Persons, situations, and virtue ethics. *Nous 32*(4): 504–530.

Doris, J. M. 2022. *Character trouble: Undisciplined Essays on Moral Agency and Personality*. Oxford University Press.

Fischer, P., J. I. Krueger, T. Greitemeyer, C. Vogrincic, A. Kastenmüller, D. Frey, … & M. Kainbacher. 2011. The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin 137*(4): 517.

Kamtekar, R. 2004. Situationism and virtue ethics on the content of our character. *Ethics 114*(3): 458–491.

Latane, B., & J. M. Darley. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts.

Merritt, M. 2000. Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice 3*(4): 365–383.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. HarperCollins.

Stichter, M. 2020. Learning from failure: Shame and emotion regulation in virtue as skill. *Ethical Theory and Moral Practice 23*(2): 341–354.

PART III
Agency and Moral Responsibility

So far, we have been focused on the "inside" of moral behavior and judgment, asking questions about the inner workings of a person's psychology. What causes people to act and what motivates people to act morally? How do these motives develop? If desires are at the heart of moral motivation, what do people desire? Are moral judgments inherently motivating? Do our brains produce morally biased judgments? Do people have virtuous character traits? We can also take an "outside" point of view on moral motivation. When we see someone else behaving badly we can ask whether it makes sense to hold them responsible for what they did. Should they be blamed or punished? If someone does something we deem to be good, should we praise them? Can we blame Mean Mary for running over Spot the dog if she was just born that way (with overwhelming mean desires) or if her parents trained her to be mean for their own nefarious purposes?

Of course we do often hold people responsible, and sometimes we even punish them. But it's also true that we don't always hold people responsible: we excuse people who are too young, mentally ill, or somehow incapacitated. What makes the difference between a responsible person and one who isn't responsible? Or between an action for which you are responsible and one for which you're not? Are people ever really morally responsible for what they do? The answers to these questions are often taken to depend on whether we have free will. In this part of the book we'll talk about responsibility, praise, blame, free will, and the facts about our psychology that support or challenge our ordinary ideas about moral responsibility.

DOI: 10.4324/9781003402787-12

10 The Psychology of the Responsible Agent

In the 1990s, Bernie Madoff took large amounts of money from people who trusted him with the promise that he would manage and invest it. Instead, Madoff constructed a giant Ponzi scheme – a scheme in which you use the money from new investors to pay off the older investors, rather than actually investing the money and making real profits – and defrauded people out of billions of dollars in total. Americans were angry with Madoff and blamed him for what he did. His victims sent emails to the court that tried him: "One victim called him a Bastard, with a capital B. Another labeled him 'scum' and 'a vicious animal.' Still another wrote, 'I feel like I have been economically raped'" (Keteyian 2010).

Sometimes, though, we do not want to blame people who do terrible things. Consider the outcry from human rights organizations about the fact that mentally retarded individuals could be executed for committing crimes in the United States.[1] According to Human Rights Watch,

> The United States is almost alone in the world in allowing this barbaric practice. At least 33 mentally retarded men have been executed since the United States reinstated the death penalty in 1976. Some experts estimate that as many as 10 to 15 percent of the 3,000 men and women on the nation's death rows are retarded … The mentally retarded can never meet the criteria of extraordinary blameworthiness. People with retardation are

DOI: 10.4324/9781003402787-13

> incapable of calculated, mature evil. A retarded person is simply not the same as other adults. They are childlike in many of their limitations: their ability to reason and develop skills needed to navigate in the world are permanently stunted.[2]

It seems inappropriate to be angry at a person with limited cognitive capacities, just as it seems appropriate to be angry and resentful towards Madoff.

Whatever you think about these cases in particular, it is clear that we do hold some people responsible for their immoral actions and that our attributions of responsibility vary depending on the psychological characteristics of the agent. There are many examples of this from ordinary life. You would likely blame your roommate for punching you in the face or for taking your sandwich, but you wouldn't blame a two-year-old child for hitting you or for taking something that doesn't belong to her. We don't tend to hold someone who is experiencing a nervous breakdown as responsible for bad behavior as someone who is mentally healthy. We excuse people for reasons of temporary insanity. Psychological factors shape our attributions of moral responsibility, praise, and blame.

Holding People Responsible

What distinguishes the psychology of the person we hold responsible from the psychology of the person we excuse? In other words, what psychological states must an agent have for it to make sense to praise or blame her for what she does? This is not the only question we could ask about moral responsibility, but it is an excellent place to start in a book about moral psychology. And we already have some possible answers: mental health and psychological maturity are two candidates that came up in the opening examples. Notice that this is not a question simply about what people actually do in their praising and blaming behavior; it is a question about what it makes sense to do, what is morally or normatively appropriate. It might be that we sometimes, in fact, hold people responsible unfairly, as we once did with mentally retarded people according to Human Rights Watch. What we're looking for here is a characterization of when and why it *makes sense* to hold an agent responsible.

Now, you might think that this gets things backwards and that to establish whether it makes sense to hold people responsible at all we need to figure out, first, if they are metaphysically responsible or transcendentally free, or something like that. In other words, you might think that we have to establish whether determinism is true, and if it is, whether true freedom of the will is compatible with determinism, before we can get anywhere on the topic of responsibility. Many philosophers reject this view, however. Some think that the question about responsibility is an inherently normative (or moral) question and some think that we need to figure out what responsibility looks like before we can ask about the metaphysical conditions for it. In this section we'll consider these arguments and then proceed to consider the psychological conditions of responsibility along these lines. We will leave further considerations about determinism for the next chapter. (If you really can't put the free will question aside for now, you can skip ahead to Chapter 11 and come back to this one.)

Peter Strawson is more responsible than anyone (no pun intended) for drawing philosophers' attention to the importance of *holding* people responsible. In his famous 1962 paper "Freedom and Resentment," Strawson (2008) tries to justify our practice of *holding people responsible* without relying on an independent theory of the conditions under which people are responsible. Strawson thinks that our practice is justified because once you understand our practice of holding people responsible, you will also see that (first) it has its own norms that have nothing to do with the metaphysics of anything and (second) that it's really, really important to us and to how we relate to each other. To see Strawson's point, we need to explore his idea that our ordinary relationships with people are structured by what he calls the reactive attitudes.

The reactive attitudes are emotions that we feel toward other people insofar as we regard them as persons capable of meeting our expectations, rather than as objects that will just do what they do. These emotions are "reactions to the quality of others' wills toward us, as manifested in their behavior: to their good or ill will or indifference or lack of concern" (Strawson 2008: 15). Reactive attitudes include the kind of moral resentment and indignation we feel toward someone like Bernie Madoff; they also include positive emotions such as gratitude that we feel toward people who do good things and emotions like guilt, shame and pride that we feel about our own actions. We do not have reactive attitudes toward everyone whose actions affect us. When we do not take these actions to be representative of the quality of the other person's attitude toward us in an important way, we instead respond to them with "objective attitudes." When we take an objective attitude toward another we see them as something to be managed, handled, cured, trained or avoided; we do not see them as someone to be brought into conversation, reasoned with or engaged. If your roommate steals your sandwich from the counter, you could be resentful. If your dog does it, you'll think that you need to step up your training.

Strawson asks us to think about the cases in which our ordinary blaming attitudes toward bad behavior are mitigated. If you think about the simple case of your roommate stealing your sandwich, you can see what kinds of changes to the situation would reduce your resentment. First, if your roommate didn't know that the sandwich was yours and threw it in the garbage because he was trying to tidy up, you would probably feel less resentful. Second, if your roommate just found out her mother died and was out of her mind with grief, you would probably think that "she's not really herself" and feel less resentful than usual. Finally, if your roommate were a dog, a child or a cognitively impaired person who doesn't really understand what stealing is, you would not likely feel resentment at all (though you may be quite annoyed). In all these cases, what the mitigating circumstances reveal is that the action of stealing the sandwich doesn't mean that your roommate disrespects you, doesn't care about you, or wants to hurt you. Given this, it's not appropriate to hold her responsible in a way that entails blame and the possibility of punitive sanctions.

Strawson then argues that the question about determinism is this: “Could, or should, the acceptance of the determinist thesis lead us always to look on everyone exclusively in this way (in the objective way)?” (2008: 12). In other words, if we believe that human actions are determined – that is, causally entailed by the pre-existing facts together with the laws of nature – should we regard everyone in the way that we regard pets, children, the mentally ill, and the cognitively impaired? Strawson thinks we cannot and should not attempt to change our attitudes in this way, for two reasons. First, we already have reasons for adopting the objective attitude toward some people in some circumstances and these reasons have nothing to do with the truth of determinism.[3] Rather, they have to do with the factors we mentioned above (whether the person knows what she’s doing, whether she was herself when she did it, and so on). Second, what we would give up by taking the objective attitude toward everyone would not be worth whatever would be gained by it.[4]

To see what would be lost in giving up the reactive attitudes, it will help to think about Strawson’s picture in more detail. R. Jay Wallace elaborates Strawson’s ideas by pointing out that what is crucial to the attitudes we take toward those we hold responsible – resentment, indignation, and guilt, according to Wallace – is that these attitudes are a response to violated expectations. We expect people to behave in certain ways and when they fail, we are resentful or indignant. When we fail our own expectations we feel guilty. With this refinement of the picture in mind, think about what life would be like if we took an objective attitude toward everyone. What would happen to friendship, for example, if we refused to hold our friends to any expectations? Imagine a friendship between you and me in which neither of us expects the other to be honest or loyal or helpful; when you are mean to me, I’m not angry, and when I break my promise to you, you might feel sad but you do not resent me. This sounds very much like the relationship I had with my goldfish, though even my goldfish might have resented it when I forgot to feed him.

The above argument gives us a reason to look to psychological conditions of responsible agency, since it is obviously very important to us to distinguish responsible people from “objects.”[5] So far, we have seen some paradigm examples of conditions that undermine attributions of responsibility, but we have not yet arrived at a definitive characterization of the psychology of a responsible agent. There are many theories to consider, which can be divided into “Self-expression” theories and “Reasons-Responsiveness” theories.[6] According to the former, an agent is responsible for her actions when those actions express her real, true, or deep self. According to the latter, an agent is responsible for her actions when those actions result from her rational capacities.

Methodology

Before we turn to examine these two theories, it’s worth pausing to ask what question we are really asking and what the standards are for a good answer.

We have already seen from our discussion of Strawson that there are two questions we might ask:

- Under what conditions do we, in fact, hold people responsible? This is a descriptive question about the nature of our practice of attributing responsibility.
- Under what conditions does it make sense (or is it rational) to hold people responsible? This is a normative question about the *legitimacy* of our practice.

For many philosophers (including Strawson on some interpretations), these two questions are importantly related. This is because of a background view about the methodology for defending answers to the question about the legitimacy of our practice (the second question). According to this background view, we vindicate our practice by describing it in a way that makes sense. This methodology – *reflective equilibrium* – is very common in ethics (as we have already seen in previous chapters). Very briefly, the idea is that you defend normative theories by showing that they are the best way of systematizing our ethical principles, considered judgments (sometimes called "intuitions") about cases, and our background philosophical and scientific theories.[7] Following this method, we defend a *normative* theory of responsibility (that is, a theory that tells us when it is fair or reasonable to hold others responsible) by demonstrating that the theory best systematizes our considered judgments about cases of various kinds along with the relevant background theories.

Reflective equilibrium does not imply that all of our considered judgments are correct or that the best theory is the one that preserves the greatest number of them. Some of the judgments we make about cases may be in conflict with other judgments, and some judgments might have implications that conflict with other things we believe very strongly. For example, consider someone who starts to think about the topic of responsibility who is inclined to make two judgments about cases: first, that Harry and Susan who each donated a kidney to a total stranger (see Chapter 3) do not get any real credit or praise for this because they were just acting on sympathetic animal instincts and, second, that Bernie Madoff is the scum of the earth, deserving of all the blame we can heap upon him because his greedy decisions hurt people tremendously. On the face of it, these judgments are in tension with each other. The theorist engaged in reflective equilibrium has at least two options. First, she can point to some relevant difference between the two cases. Perhaps, albeit implausibly, Harry and Susan didn't ever think about their options but just acted from the gut without thinking it through, whereas Madoff acted in a much more calculating fashion. Or, second, the theorist could reject one of these judgments, because it is at odds with too many of the other things she thinks about responsibility or about the way the world works. For instance, she might decide to reject her judgment that Susan and Harry are not responsible, because she realizes that even if they were motivated by sympathy, they also did what they did for the reason that it benefitted others.

This example features a rather obvious inconsistency; it will become apparent later in this chapter and the next that some inconsistencies are much more difficult to notice and resolve. The important point for now is that reflective equilibrium does not require us to preserve all of our judgments. Some of them may have to be given up when we arrive at the best theory. What this means is that there may be a gap between the best theory and what we happen to think. Reflective equilibrium "cleans up" our intuitions about cases, revising and pruning as necessary to reach an equilibrium point. The bigger the gap, the more we can complain that the theory of responsibility we have been offered isn't really a theory of what we mean by responsibility. It's good to keep in mind, though, that because our various judgments and beliefs are not perfectly consistent, every theory will create some gap.

We're now ready to consider the two different theories of the psychological conditions for responsibility. The philosophers who defend these theories often appeal to examples in order to show that their theories make the *best* sense of our current practices and views about responsibility, praise and blame. Given the method of reflective equilibrium, if they are successful, they will have shown not just how we tend to attribute responsibility, but also why it makes sense to do it in this way.

Self-expression Theories

One important difference between Bernie Madoff and the roommate who steals your sandwich because she's blinded by grief is that Bernie Madoff's actions seem to express who he really is (someone greedy and ruthless enough to deceive people and make a lot of people poorer in order to become wealthy himself), while the roommate's actions do not. For this reason, Madoff seems more responsible for what he's done than the distraught roommate. Similarly, someone who dances on the table because he is under hypnosis seems less responsible for breaking your table than someone who does it because he thinks it would be fun. These insights have led some philosophers to think that the key to responsible agency has to do with acting on the beliefs and desires that are truly yours, as opposed to acting on beliefs and desires that are alien or foreign to the "real you."

Harry Frankfurt developed this sort of view in his paper "Freedom of the Will and the Concept of a Person" (1971). Frankfurt puts the point in terms of free will, rather than responsibility, but it is clear that free will *in the sense that is relevant to responsibility* is what really matters to him. Frankfurt's basic picture is that we act freely when we act on our own desires, and we have a free will when the desires that we act on are the ones we want to be acting on. In other words, free *action* is doing what we want and free *will* is doing what we want to want. Here Frankfurt is invoking "second-order desires" to explain what it is to be a person. (Note that *person* here, as in a lot of moral philosophy, is a normative category, as opposed to *human being*, which is a biological category. For Frankfurt, "person" is the word we use to describe those human beings who are capable of responsible action.) Persons have desires about which of their desires they want to move them. Your desires move you around, and if all you have is those desires, you're more like an animal than a person (you are what Frankfurt calls a "wanton"[8]). What you *want to want* determines what you, the person, "really" want as distinct from what is merely a force inside your head making you do things.

You, most likely, are a person. To see this, think about some time that you've hoped to change yourself in some way. For example, if you have ever wanted to quit smoking or to spend less time playing video games, then you have had a second-order desire (you wanted not to want a cigarette, for example) and that makes you a person in the relevant sense, according to Frankfurt. Or, think about your long-term goals (such as earning a degree or finding a satisfying job), which likely incline you to have desires about your short-term, first-order desires: you want to act on your desires to study and not on your desires to go out drinking every night, for example. Human beings who have no second-order volitions (human beings who are not persons in this sense) can certainly do things – they do have desires that they can act on – but they do not have any second-order volitions; there is no true self that they want to be.

How does this distinction between persons and non-persons help us understand the psychological conditions of responsibility? We could say that when a person is moved by the desire that he wills to move him, then he is responsible for what he did and deserving of praise and blame. In this way of thinking, people deserve praise or blame for the actions that stem from who they really are, their true or real selves. People do not deserve praise or blame for the actions that stem from desires that are imposed on them from the outside and that they do not want to have (as in the case of hypnotism). And animals or wanton human beings who just act on their desires, with no concern one way or another about what these desires are, do not really warrant praise or blame at all.

According to Frankfurt, I am most fully myself, and therefore responsible for what I do, when I do what I want to do and what I want to do is what I want to want to do. For a person to be free and responsible is for them to have the will they want (Frankfurt 1971: 15). There is surely something compelling about this theory, but some serious objections have also been raised. The main objection can be put this way: What's so great about the *second order?* Why think that who you really are is identified with your desires about your desires and not something else; why not your third-order desires, for instance? Talking about desires in the abstract makes the point kind of obscure, so let's think about this objection in terms of an example. Consider a person who suffers from anorexia nervosa, who has some (first-order) desires to eat and some first-order desires to be thin, but a very definite and unhealthy second-order desire to act on the desire to lose weight and never on her desires to eat.[9] In the view we are considering, it seems that we have to say that the person with anorexia is most truly herself, most responsible, most deserving of praise and blame, when she acts on her desire to lose weight, because this is the desire she wants to be effective. But now it seems like something has gone wrong. The problem is that our second-order desires can be the result of illness or external forces, in just the same way that our first-order desires can. Anorexia nervosa is a serious disease that affects second-order desires. The case shows that second-order desires can be just as alien to who we really are as first-order desires. Further, a person who is recovering from anorexia might have a third-order desire to be rid of her second-order desire to act only on her first-order desire to lose weight. Why shouldn't we think that this third-order desire is more representative of who she really is?

Frankfurt responded to this objection by fine-tuning the characterization of the higher-order attitude that represents our true selves. Ultimately, he argued that we act freely, in a way deserving of praise and blame, when we act on desires that we wholeheartedly endorse (Frankfurt 1988). The hope is that wholehearted endorsement would block the move to further orders of desire, because to endorse something with your whole heart settles things for you. It is a decisive act of commitment to being a certain way.

According to Nomy Arpaly and Tim Schroeder (1999), however, wholehearted endorsement doesn't fix all the problems with Frankfurt's theory. They think this because "endorsement" is a cognitive response of the rational or judging part of the self and this implies, problematically in their view, that we are not responsible or praiseworthy when we act on desires that might be very good parts of us even though we don't judge them to be. To see their point, consider the case of Mark Twain's character Huckleberry Finn. Huck Finn is a good southern boy who was raised to think that some people rightly own slaves and that the laws upholding the institution of slavery are just. Nevertheless, when Huck finds himself becoming friends with Jim, an escaped slave, he finds that he cannot turn Jim in to the authorities even though he believes that is the right thing to do. Huck sees himself as a bad boy who cannot bring himself to do the right thing as he contemplates letting Jim escape:

> It made me shiver. And I about made up my mind to pray, and see if I couldn't try to quit being the kind of a boy I was and be better. So I kneeled down. But the words wouldn't come. Why wouldn't they? It warn't no use to try and hide it from Him. Nor from ME, neither. I knowed very well why they wouldn't come. It was because my heart warn't right; it was because I warn't square; it was because I was playing double. I was letting ON to give up sin, but away inside of me I was holding on to the biggest one of all.
>
> (Twain 1994/1884: 161)

We feel differently, though. For readers of Twain's novel, Huck is doing the right thing by not turning Jim in, and we are relieved when he decides, "All right, then, I'll GO to hell" and tears up the letter reporting Jim to his owner. Huck's compassionate, friendship-based desires are leading him in the right direction, against his reason or at least against what he believes he should be doing. Arpaly and Schroeder call this a case of "inverse akrasia" because it is a case in which Finn acts *rightly* against his better judgment, as opposed to regular cases of akrasia (weakness of will) in which people act *wrongly* against their better judgment.

Arpaly and Schroeder think that cases like this – cases in which we act morally against our better judgment – cause a problem for Frankfurt, insofar as "wholehearted endorsement" refers to a judgment. The problems stems from what they take to be the basic intuition that Huck is praiseworthy for not turning Jim in, precisely because he does what his moral desires urge him to do. In her book, *Unprincipled Virtue* (2003), Arpaly argues that the psychological condition that grounds attributions of responsibility is the quality of a person's will. A person with a good will deserves praise, and a person with bad will deserves blame. Wills are complicated things that are comprised of desires and feelings: being motivated to do good and feeling good about doing it.[10] Given everything that is at stake, Huck has a better will than a similar boy who would listen to his judgment and turn in his friend for punishment. According to Arpaly, Huck is therefore praiseworthy for what he did. Chandra Sripada (2016) argues for a similar view about moral responsibility, which he calls the "deep self" theory. According to this theory, your real self is defined in terms of what you care about where a care is not a simple judgment or feeling but "a complex syndrome of motivational, commitmental, evaluative, and affective dispositions" (Sripada 2016: 1211).

Though the theories we have discussed in this section disagree about what is the most important part of the self, they agree that it is helpful to think about responsibility in terms of the self. We could put the basic insight this way: you should be praised or blamed for the actions that stem from who you really are, deep down, or from your character, since actions that do not result from your real self or character must have some cause that is external to you. This was the view of David Hume:

> Actions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil. The action itself may be blameable; it may be contrary to all the rules of morality and religion: but the person is not responsible for it; and as it proceeded from nothing in him, that is durable or constant, and leaves nothing of that nature behind it,'tis impossible he can, upon its account, become the object of punishment or vengeance.
>
> (2000/1739: 411)

We can see the debate between the two theories we've been discussing as a debate about what a person's true character or real self is. Frankfurt's wholehearted endorsement view, Arpaly and Schroeder's whole-self view, and Sripada's deep self view are different approaches to identifying your *real self*, that is, the person you really are. Should the self that is responsible be identified with desire, wholehearted endorsement, or our deepest cares and emotional concerns? A compelling answer to this question might help sort the cases in the right way and reach a good equilibrium about responsibility.

On the other hand, one might think that talking about the real self is the wrong approach to thinking about the psychological conditions for responsibility. After all, what your "self" is like is largely a product of your upbringing and your culture, so why should you be held responsible for it? Thoughts like these have led to the development of a different approach to moral responsibility.

Reasons-Responsiveness Theories

What if our real self or our character (what we want, what we endorse, what our emotions are and so forth) is beyond our control? If it is, and if our real self determines how we act, then it looks like we aren't really responsible for what we do after all. Sure, Bernie Madoff defrauded hundreds of people out of millions of dollars on purpose and because he was a bad guy, but if he didn't choose to be a bad guy in the first place, how does being a bad guy make him responsible? Maybe we should look elsewhere for the psychological conditions of responsibility.

Obviously, we do not have control over the past. When we decide what to do now, we cannot change the influences of our upbringing and culture so that we have different options open to us now. But we do have *some* kind of control. Right now, you could lift your arm up and wave your hand around. Or, you could not do that. Now that I've put the idea in your head, you could do it or not do it – however you were raised, it's up to you. There are different kinds of control we might have over our actions, some of which we are more likely to actually have than others. The philosophers we'll discuss in this section think that the most important kind of control has to do with our capacities to recognize, grasp, and act for reasons. Such theories are often called "reasons-responsiveness theories" and that's what we'll call them here.

In short, as Dana Nelkin puts it, the idea behind reasons-responsiveness theories is that "people are not responsible for their actions when they lack either the capacity to grasp reasons for acting (or not acting, as the case may be) or the capacity to translate those reasons into action (or omission)" (2008: 498). Intuitively, reasons-responsiveness theories will say that the difference between the roommate who steals your sandwich and the dog who steals your sandwich is that the roommate knows that there are reasons not to take your food and is capable of refraining from stealing the sandwich and satisfying their hunger in some other way while the dog has no such capacities. Of course, the dog may be well trained not to grab food from the tops of counters, but it isn't grasping or considering "do not take things from others without their permission" as a reason to act in a certain way. Notice that if the dog takes the sandwich anyway, we're more likely to blame the trainer (or the smell of bacon) than we are to blame the dog. Some may decide to punish the dog as a way to correct the behavior in future, but it doesn't really make sense to think that the dog *deserves* the punishment.

To understand reasons-responsiveness theories more fully, it's useful to introduce a distinction that will illuminate the different types of control we might have over our actions. John Martin Fischer and Mark Ravizza (1998) distinguish *regulative control* and *guidance control*. Regulative control requires genuine alternative possibilities open to the agent so that she could really do something other than what she actually ends up doing. Guidance control is what you have when it is your decisions that make you take the option that you do take even though, if your decisions themselves are caused by something outside of you, you were not free to decide otherwise (and hence you did not really have alternative possibilities open to you). To put it in a simple way, regulative control is the ability to choose and act otherwise than you do, guidance control is the ability to act on your own choices for your own reasons.

The importance of all this for our purposes, according to Fischer, is that guidance control is all that is required for moral responsibility (Fischer, Kane, Pereboom, & Vargas 2007). This is a good thing for those of us who want to hold on to responsibility, because guidance control is (according to Fischer) probably the only kind of control we have. Importantly, even if we do not have regulative control, this doesn't matter because we do have guidance control, which is sufficient for moral responsibility. What exactly does guidance control amount to?

Defenders of the reasons-responsiveness theory think that the key to guidance control – and hence moral responsibility – is a moderately reasons-responsive mechanism that is the agent's own. To be responsive to reasons is just to understand what considerations there are in favor of doing this or that and to be able to act for the right reasons. So, as Nelkin (2008) puts it, the relevant competencies for moral responsibility include both cognitive and volitional capacities. The idea that these capacities must be *moderately* well tuned is important, because if you define the relevant capacities at either extreme, you get into trouble. To say that a mechanism is *moderately* responsive is to say that it might not respond to every good reason every single time, but it does respond by and large to the patterns of reasons that confront it. Why do the extremes cause trouble?

On one hand, a mechanism that is *strongly* reasons-responsive would be one that never fails to act for the good reasons that there are. Consider the case of Befuddled Bob, who is deciding whether to follow his friends to the bank in order to rob it or to ditch them. On the assumption that there are several excellent reasons not to follow his friends in their dumb plan, Bob would not count as *strongly* reasons-responsive unless he decided to ditch his friends. If strong reasons-responsiveness were required for responsibility, this would mean that Bob could never be responsible for helping his friends to rob the bank, because this option isn't supported by the best reasons. Indeed, in this way of thinking, Bob could not be responsible for doing anything that wasn't a good idea! But this seems like the wrong thing to say: the idea behind the reasons-responsiveness view is that we should be held responsible for doing something dumb when we are capable of seeing the good reasons not to do it, but we ignore them.

On the other hand, we could weaken reasons-responsiveness too much so that someone who responds to the most ridiculous kinds of reasons counts as reasons-responsive. For example, if Befuddled Bob were responsive only to suggestions from people with tattoos, there's a sense in which he would be responsive to reasons, but it would be a sense that makes him kind of crazy, rather than a paradigmatically responsible agent. Reasons-responsiveness theories need to steer a path between strong and weak reasons-responsiveness to something moderate. A person who is moderately responsive to reasons is someone who, by and large, is guided by elements of the coherent patterns of reasons that confront us when we make decisions, though she may not always be guided by every reason in the pattern every time. As David Brink and Dana Nelkin put it "Where there is sufficient reason for the agent to act, she regularly recognizes the reason and conforms her behavior to it" (2013: 294).[11]

According to reasons-responsiveness theories, then, to have guidance control a person must have a psychological mechanism that is moderately responsive to the reasons that there are for acting one way or another. When this mechanism guides your actions, you are responsible for what you do. When you behave independently of this mechanism (as you might, say, if you're sleepwalking), you are not responsible for what you do. A person in whom this mechanism is seriously defective or non-existent is not a responsible agent. In this way of thinking about responsibility, Madoff does seem to be responsible for defrauding seniors out of their retirement money: he acted for selfish reasons that seemed like the best reasons to him.

So, a person who is responsive to reasons has a particular set of psychological capacities. We might wonder how these capacities develop in human beings and whether this matters to responsibility. Consider that in order to grasp reasons at all, we need to be attuned to them in some way; that is, we need to see that the fact that a statement is true is a good reason to believe it, and the fact that a course of action is morally required is a good reason to do it. Our thinking about reasons has to be hooked up in the right way to what normative reasons are, and this is something we learn as we develop from unreasonable children into rational adults. Does the fact that these capacities develop have any implications for reason-responsiveness theories?

Susan Wolf has an interesting answer to this question in her paper "Asymmetrical Freedom" (1980). The asymmetry that she notices is between our judgments of responsibility for bad actions and our judgments of responsibility for good actions. Wolf notices that we have a different standard for freedom when we praise people for doing good things than we do when we blame people for doing bad things. "Here I stand. I can do no other," Martin Luther is supposed to have said when he refused to recant his radical religious writings. This kind of integrity is paradigmatically praiseworthy, even though it might really be true that Martin Luther really couldn't do anything other than what he did, given what his conscience told him. On the other hand, a person who has been brainwashed into thinking that he can do no other than turn to crime is someone we take to be less blameworthy than the person who chooses crime even though he didn't have to. This asymmetry leads Wolf to conclude that it is not whether we could do other than what we were determined to do that's important, but, rather, whether we were determined *in the right way*. Further, being determined in the right way has to do with whether our capacity to recognize and act for reasons tracks the truth about the reasons that there actually are. On the assumption that Martin Luther was well brought up to distinguish between right and wrong, the fact that he can only do what is right (that he is psychologically determined to do what is right) does not impugn his agency or undermine our holding him responsible.

Wolf's argument draws our attention to the fact that "reasons-responsiveness" develops in human beings as a result of explicit or implicit education. If a person is poorly raised, they might be free to act immorally in ways that the rest of us are not, but this doesn't make the person more responsible. It makes the person a sad case. Once we recognize that reasons-responsiveness is a psychological capacity that develops in the way that other capacities do, we can begin to wonder whether we really do have this capacity and whether it is as effective as we might hope. We'll turn to these questions in the next section.

Are We Responsible? Challenges from Psychology

According to reasons-responsiveness theories, we are responsible for our actions when they result from the capacities that allow us to be guided by reasons. These theories do not necessarily posit a "real self" that is the source of free actions, but they do assume that we have certain rational capacities – the capacities to grasp and act for reasons.[12] Do we really have these capacities? There is some cause for skepticism.

In Chapter 7, we discussed moral dumbfounding – the phenomenon of people sticking to their moral judgments even when they have no reasons for them – which might look like evidence that our capacity to be guided by reasons is not very strong. I argued that this evidence doesn't cause too much trouble for Kantianism, but it's worth looking at more evidence that we're not as rational as we think we are in the context of reason-responsiveness theories of moral responsibility. There is a substantial body of research that claims to show that we often do not grasp reasons and act for those reasons. For example, a number of studies suggest that our choices of consumer products are not made for the reasons we think they are. These studies also tend to show that we confabulate reasons after we choose so that our choices *seem* to have been made for considerations we endorse as reasons, even though they really weren't. For example, in one such study, people were asked to choose among four pairs of stockings that were (unbeknownst to the shoppers) exactly the same. People tended to choose the stockings on the right, but they explained their choice by referring to the better quality of the stockings they chose (Wilson & Nisbett 1978).

Other, more elaborate, studies show that analyzing our reasons tends to change our attitudes toward political candidates, our beliefs about whether our romantic relationships will last, and our judgments about how much we like different posters (Wilson, Kraft, & Dunn 1989; Wilson & Kraft 1993; Wilson, Lisle, Schooler, Hodges, Klaaren, & LaFleur 1993). In this last study, participants were asked to evaluate two types of posters: reproductions of impressionist paintings and humorous posters, such as a photograph of a kitten perched on a rope with the caption, "Gimme a Break" (Wilson et al. 1993). All of the participants were asked to rate how much they liked each poster and then allowed to choose a poster to take home. The reflectors were instructed to write down their reasons for feeling as they did about the posters before giving their ratings, while the controls did a cognitive task not related to reflecting on reasons (a filler task). The results of the study were that the reflectors rated the humorous posters significantly higher than the controls did and were much more likely to take these humorous posters home.[13] A few weeks later, the researchers telephoned all the participants and asked them several questions about the posters they had chosen (how much they liked them, whether they still had them, whether they had hung them up on their walls). The reflectors were less satisfied with their posters than the people who did not reflect on their reasons before choosing a poster.

Apparently, what happens is that people don't really have reasons for many of their attitudes, beliefs, and judgments (such as their poster preferences), so when they are asked to analyze their reasons, they just make up something that's easy to think of or that they believe will make sense to other people – that is, they confabulate. These confabulations lead people away from the attitudes, beliefs, and judgments they made before they started thinking about it (such as the preference for the impressionist poster). There's nothing necessarily wrong with this – analyzing your reasons might improve your beliefs, after all. The point for our purposes is that people think their beliefs and judgments are *based on* their reasons, but this is mistaken if the reasons they confabulate create new attitudes that they didn't have before they thought about it (such as the preference for the kitten poster).

Taken as a whole, the psychological evidence suggests that we do not often know why we do what we do, and when we look for reasons, we look for easy to find considerations that we believe will make sense to people as explanations for what we have done, whether or not those were really our reasons for acting.[14] In other words, we often engage in *post hoc rationalization* of our actions, rather than grasping the reasons that we really do have, deciding which ones to act on and then deliberately acting for those reasons. Further, when we act, we are frequently caused to act by factors that we would reject as reasons for action if we thought about it. If this is how we are, then the picture of us as competent rational agents, deliberating about our reasons and then acting on the results of our deliberation, seems to be a bit tarnished.

But how tarnished, really? Is it true that we do not have the rational capacity to reflect on, endorse and act for reasons that is required by responsibility? The evidence we have surveyed is evidence that we don't use this capacity all the time, but it isn't evidence that we don't have it. If we *sometimes* act for the reasons we think we do – if we sometimes reflect on, endorse and act for the reasons we recognize as reasons, even though not always – this will be enough to show that we are at least moderately responsive to reasons, which is enough for responsibility, according to the reasons-responsiveness theory.

Importantly, the psychologists who have done the research I have presented as evidence against the effectiveness of our rational capacities do not tend to endorse any strong claim about our lacking these capacities altogether. As their research has continued, they have found that there are some variables that seem to make us better at knowing our reasons. For instance, the more knowledgeable we are about something, the more we understand our own reasons for making choices with respect to that thing (Halberstadt & Wilson 2008). Someone who was an expert in stockings would probably have seen that the stockings were identical (as a few people in the actual study did) and would not have fabricated reasons for choosing one over another. Someone who was able to articulate what it is about impressionist art that makes it beautiful might have chosen the art poster rather than the kitten poster after reflecting on their reasons for preferring one to the other. Knowledge is not all powerful, of course, but it does sometimes help, which is evidence that we should not be so pessimistic as to think we never do things for the reasons we think we do.

To illustrate these ideas, consider this former participant in the Milgram obedience experiment discussed in Chapter 9. Jacob (not his real name) describes to the science writer Lauren Slater how at around the same time as he participated in Milgram's study, he was struggling to accept his homosexuality. Jacob was one of the people who shocked the "learner" to the highest degree; he was 100 percent compliant. The experience of being in the experiment had a profound effect on him:

> The experiments … caused me to reevaluate my life. They caused me to confront my own compliance and really struggle with it. I began to see closeted homosexuality, which is just another form of compliance, as a moral issue. I came out. I saw how essential it was to develop a strong moral center. I felt my own moral weakness and I was appalled, so I went to the ethical gym, if you see what I mean.
>
> (Slater 2004: 59)

Jacob may not have acted for reasons he could endorse in the experiment. But his behavior in the experiment gave him reasons that guided his actions for the rest of his life. As we saw, philosophers who think being guided by a reasons-responsive mechanism is necessary for moral responsibility do not think that people are always perfectly responsive to the reasons that there are. Rather, they think that this is a capacity we have, which we sometimes act on and sometimes don't. Similarly, the rational capacity to acknowledge and act for the reasons we think we have need not be a capacity that we use all the time in order for us to be responsible agents.

There is also empirical evidence from psychology and neuroscience that the capacities needed for responsibility correspond to brain functions that normally develop during a person's life. For example, Katrina Sifferd and her colleagues argue that reasons-responsive capacities are a part of "executive function," which is the name for several processes, including "attention, (considered) recognition, memory, decision making, the planning of intentional actions, and the inhibition of actions" (Sifferd & Hirstein 2013: 132; see also Hirstein, Sifferd, & Fagan 2018). It's not hard to see how these capacities overlap with the capacities to recognize and act for reasons. To grasp reasons for acting one way or another, we need to have the capacity for attending to our situation. To act for those reasons, we need the capacity to decide to do so and plan our actions accordingly. And to act successfully on our reasons despite competing temptations, we need the capacity to inhibit tendencies to act that are contrary to our reasons. If the reasons-responsive mechanism is just another name for executive function, then we certainly have it, though to varying degrees.

Empirical research does not prove that we are not rationally competent creatures. (Indeed, some research points us to where to look for the relevant capacities in the brain.) It does not prove that we cannot shape our behavior over the long term in the light of considerations that we consciously take to be reasons that favor moral behavior over immoral behavior. It therefore does not undermine moral responsibility, if reasons-responsiveness theories are correct. There are real psychological differences between people who are hypnotized to do things and people who choose to do them, and between people who are incapable of understanding what a moral reason is and people who just ignore moral reasons out of selfishness or malice. These differences seem to make a big difference to our practice of praising, blaming, and holding people responsible.

A Case Study: Implicit Bias

As we've already noticed, responsibility is not an all or nothing matter. For example, according to the reasons-responsiveness theory, your capacities can be more or less engaged, more or less under strain, and more or less bypassed because of features of the situation. Recall from our discussion of the empirical challenge to virtue ethics in Chapter 9 that our actions are often influenced by situational factors that we would not endorse as reasons for acting. These examples – like the people who fail to help a needy person when there are passive bystanders around – provide some evidence that sometimes our capacity to respond to reasons is not very sensitive. One significant example of the situation subverting our rational capacities that is often discussed in the literature on moral responsibility is the case of implicit bias. This case is worth discussing because it raises some interesting practical questions.

Discussions of implicit bias are ubiquitous these days in universities, businesses, and the media. Implicit bias refers to prejudiced attitudes that are automatic, often (though not necessarily) inconsistent with the explicitly endorsed attitudes of the person who has them, but still capable of influencing their judgments, decisions, and actions. Someone who clutches her purse and crosses the street whenever she sees a Black man, but claims up and down that she is not racist, probably has some *implicit* racist bias. The problem that implicit bias raises for moral responsibility, is that it's hard to see how we can be responsible for something that we don't know is there and that contradicts our explicit judgments about equality and fair treatment.

To see the problem more sharply, let's consider two cases from Noel Dominguez (2020: 163).

Toxic Environment – Dawn grows up near an illegal but well-hidden toxic gas dumping site, and as a result has a number of symptoms she wouldn't have if she'd grown up somewhere else. One such symptom is that she can often be condescending and impatient with others without meaning to or having any reason to. In fact, she often doesn't notice she's doing this until she is told how inappropriately she has been acting, at which point she feels confused and ashamed. She genuinely feels bad about what she does and wishes she would not be so dismissive, but has trouble exercising any direct influence over this reaction.

Toxic Social Environment – Dave grows up near an immoral but well-hidden toxic racist social environment, and as a result has a number of symptoms he wouldn't have if he'd grown up somewhere else. One such symptom is that he can often be condescending and impatient with minorities without meaning to or having any reason to. In fact, he often doesn't notice he's been doing this until he is told how inappropriately he has been acting, at which point he feels confused and ashamed. He genuinely feels bad about what he does and wishes he would not be so dismissive, but has trouble exercising any direct influence over this reaction.

Perhaps you can already anticipate the problems for moral responsibility raised by implicit bias. In short, Dawn seems like a good case of someone who isn't responsible and should not be blamed for her undesirable attitudes, whereas Dave seems like a case in which we are quite tempted to say he does have at least some responsibility for acting on his implicit bias. In part, of course, we're tempted to hold Dave responsible because implicit bias has really bad consequences for those who are the target of it – in hiring practices, education, health care, the criminal justice system, and more.[15] As Dominguez puts the problem: "If Dave really is responsible for his actions here, then we'd either need an explanation of how his case differs from Dawn's, or an explanation of why taking Dawn to be responsible isn't so far-fetched" (2020: 163).

Let's think about what the various theories we have looked at should say about implicit bias. Self-expression theories like Frankfurt's that include explicit endorsement or wholeheartedness do not allow us to hold Dave responsible. Dave is just as internally conflicted as Dawn – neither wholeheartedly endorses their negative attitudes and so neither is properly held responsible for them. But self-expression theories that include other, non-rational, aspects of the self, like Arpaly's whole self theory, may have a different result. As Jules Holroyd and colleagues put it “notions of responsibility that make reference to the “real self” – the part of the self that reveals where the agent stands – may have to account for the agent’s real self as including some unendorsed implicit attitudes, as well as her endorsed explicit attitudes” (2017: 9). It seems open to the whole self theory to say that Dave is responsible, because his implicit attitudes are part of his whole self, just as Huck Finn’s anti-racist attitudes were part of his. Can whole self theories distinguish between Dawn and Dave? Or are they forced to hold both Dawn and Dave equally responsible? The answer to that question, for Arpaly, would see to depend on what it means to say that something is part of a person’s will.

At first glance, reasons-responsiveness theories do not seem well suited to distinguish between Dawn and Dave. Since neither person is aware of their implicit biases, and since both disavow those biases as reasons for discrimination, the actions that result from their implicit biases do not seem to result from their capacities to grasp and act for reasons. Rather, they seem like actions caused by forces that have completely bypassed those capacities. However, reasons-responsiveness advocates might have some sensible things to say about Dave’s responsibilities. For one thing, they can point out that once Dave does become aware of his implicit bias, he has a responsibility to do what he can to change it. Dawn would have the same responsibility, of course, but that doesn’t seem like the wrong conclusion. Going forward, if there is anything she can do to mitigate the effects of the poison on her behavior, shouldn’t she do it? Reasons-responsiveness advocates could also point to indirect control as a strategy for mitigating the effects of attitudes that work below the level of consciousness. Indirect control – sometimes called “ecological control” – aims to change behavior by changing the environment (Washington & Kelly 2016). In the context of implicit bias, indirect control could take many forms. For example, a hiring committee could make sure that all references to gender or race are removed from the stack of job applications to be reviewed and it could follow its organization’s best practices for reducing bias in hiring.

Some philosophers and psychologists, reflecting on cases like implicit bias, have come to the conclusion that we ought to revise our ordinary ideas about holding people responsible. The resulting theories are called *revisionist theories* of moral responsibility. Manuel Vargas (2009) is an advocate for revisionism. He argues that we should concern ourselves less with whether or not an individual person could have done otherwise and more with the social context of which that individual is a part.[16] On his view, we should see that it does make sense to hold people responsible as part of a social practice that aims to improve things for everyone, as long as the social environment is sufficiently supportive (Vargas 2017). Vargas is concerned with our capacities to grasp and act for reasons (and so he is often considered to be on the reasons-responsiveness “team”), but he sees attributions of responsibility as having a forward-looking purpose that aims to improve agency. This view about responsibility fits well with what certain psychologists have been arguing about implicit bias: that it’s more helpful to think of it as located in the situation, rather than in the individual mind (Murphy & Walton 2013; Payne et al. 2017).

Taking Stock

The theories of moral responsibility we have looked at in this chapter agree with Strawson that we attribute responsibility to a person when we see their actions as reflecting something important about them – their deep self, their whole self, their capacities to act for reasons, or their executive function. When a creature's actions reflect the coercion of another agent or a drug, or a lack of sleep or physical obstacle, we do not attribute responsibility to them. Both self-expression theories and reasons-responsiveness theories make sense of the different judgments we make about a variety of cases, even though they don't necessarily agree about every case. By the method of reflective equilibrium, the theory that best explains the judgments that are most central to our practice of praising, blaming, and holding responsible is the best theory. Returning to the distinction between the descriptive question and the normative question from the beginning of the chapter, we can say that the description of our practice that makes sense of the ways that we distinguish between responsible persons and non-responsible objects will also vindicate this practice. If reasons-responsiveness is the best fit, then it's right and reasonable to hold people responsible when they act in virtue of their reasons-responsive capacities.

What is key to understanding all of these theories is that they think morally responsible action is action that is *caused* in a certain way. Hence, the theories we have discussed so far – self-expression and reasons-responsiveness theories – are compatible with the claim that all of our actions are caused by some psychological mechanism or other. Because of this, these theories are called *compatibilist theories*. Self-expression theories say that we're responsible when our actions are caused by the desires, emotions, or judgments that make up our real, whole, or deep self. Reasons-responsiveness theories say we are responsible when our actions are caused by our capacities to grasp and act for reasons. One might think, however, that having the right psychological mechanism as a cause of action is not sufficient for moral responsibility. Indeed, one might think that the very fact that these are psychological *mechanisms* is a serious problem! To be morally responsible for an action, the objection goes, we have to be free, and if we are free then we are not determined to do what we do by any mechanism psychological or otherwise. We turn to this set of problems in the next chapter.

Summary

- When we hold someone responsible for an action, we regard her as a person rather than an object and we take the reactive attitudes (guilt, anger, forgiveness, etc.) to be appropriate responses to her. Taking our fellow human beings to be persons is extremely important to how we interact with each other and to our social and personal relationships.
- One way of asking when it is appropriate to hold people responsible is to ask, what is the difference between the psychology and capacities of

persons and of non-persons? Or, more specifically, what psychological capacities are necessary for appropriately holding someone responsible?

- Philosophers typically use the method of reflective equilibrium to answer these questions.
- One approach to characterizing the psychology of the responsible agent is the approach of self-expression theories.
- According to one version of self-expression theory, what is distinctive about someone acting responsibly is that they act on their second-order desires or the desires they endorse. This is Frankfurt's "wholehearted endorsement" theory.
- According to a second version, what is distinctive about someone acting responsibly has to do with their whole or deep self, including their emotional responses to their options.
- A different approach says that what is crucial for responsibility is the capacity to grasp and act for reasons. Theories that take this approach are "reasons-responsiveness" theories.
- Some research in social psychology makes it seem that we do not act for reasons that we consciously endorse as reasons; rather, we act because of situational factors and then we make up reasons to justify what we did after the fact.
- Social psychology research has not established that we have no capacity to grasp and respond to reasons, though it might show that we use this capacity less often than we think we do.
- Implicit bias raises interesting questions about moral responsibility, because it influences people's actions without their endorsement or explicit knowledge.

Study Questions

1. What would life be like if we made no distinction between responsible persons and non-responsible objects? What would change?
2. The "wholehearted endorsement" theory and the "whole self" theory offer different views about the psychological profile of action for which we are responsible and deserving of praise or blame. Think of your own example (or pair of examples) that illustrates the differences. Are there reasons for favoring one view over the other?
3. Do you think we have real selves? How do you identify yours?
4. What is the difference between "guidance control" and "regulative control"? Try to think of some examples that illuminate the distinction by showing how the two could come apart.
5. If you were on a jury trying a murder case, what kind of questions would you want to ask about the defendant's mental state in general or at the time of the crime? What could you learn that would incline you to find them "not guilty" or "not guilty by reason of insanity"?

6. Is there an important difference between Dawn and Dave? If so, what is it? If not, what should we say about who is responsible for the bad consequences of implicit bias?

Notes

1 This practice ended with the 2002 US Supreme Court decision *Atkins v. Virginia.*

2 www.hrw.org/news/2000/01/03/mentally-retarded-dont-belong-death-row. Last accessed: October 11, 2022.

3 This is the reason emphasized by Pamela Hieronymi (2020) in her subtle and fascinating interpretation of Strawson. Hieronymi interprets Strawson as saying that our practices of holding responsible and excusing are responsive to what is statistically normal and that within our practice, no general argument about determinism could give us a reason to exempt everyone with normal capacities from responsibility.

4 Strawson also says that even if we *should* give up the reactive attitudes, we would not be able to. If the facts led to the conclusion that no one is ever deserving of resentment, gratitude and so on, human beings would resist acceptance of the facts.

5 David Shoemaker (2015) develops an account of moral responsibility inspired by Strawson that recognizes more categories than "person" and "object." Shoemaker argues that agents with different kinds of capacities (including those with what he calls "marginal agency") may be responsible in different senses.

6 In the previous edition of this book, I used different labels. I have changed the labels for clarity and to be as consistent as possible with current practices.

7 See Chapters 1 and 8, and for an introduction, see Daniels (2008).

8 As in "a wanton human being." *Wanton* is ordinarily an adjective in English, meaning reckless, careless or lawless, but Frankfurt uses it in a special sense.

9 This example comes from Nomy Arpaly's (2003) insightful discussion of the case of anorexia and its implications for moral psychology.

10 In more recent joint work, Arpaly and Schroeder (2014) characterize the will in terms of intrinsic desire.

11 Fischer and Ravizza (1998) add that the mechanism in question must be "one's own"; that is, the person must *take* responsibility for her reasons-responsive mechanism; she must see it as her own and as making it appropriate for other people to hold her responsible (1998: 207–239). This qualification helps avoid some counter-examples.

12 Nahmias makes the point that *most* theories of free will – including reasons-responsiveness theories, self-expression theories, and libertarian theories – agree with the assumption that free will requires that "one's actions properly derive from decisions or intentions that one has at some point consciously considered, or at least that one would accept, as one's reasons for acting" (Nahmias 2011: 353).

13 Ninety-five percent of the controls (who did not reflect on their reasons) but only sixty-five percent of reflectors chose the art poster.

14 An important paper that helped to start this line of research is "Telling More Than We Can Know" by Nisbett and Wilson (1977). Timothy Wilson's *Strangers to Ourselves* (2002) is an accessible book that provides a balanced and engaging discussion of this research. For a discussion of the ethical implications of this research, see Tiberius (2009).

15 For a review of the effects of implicit bias in health care see FitzGerald and Hurst (2017). See Holroyd et al. (2017: 2) for many more references.

16 Other philosophers who emphasize the social dimension of responsibility include John Doris (2015) and Mich Ciurria (2019).

Further Readings

Arpaly, N. 2002. Moral worth. *The Journal of Philosophy 99*(5): 223–245.

Fischer, J. M., & M. Ravizza. 1999. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy 68*(1): 5–20.

Frankfurt, H. 1988. Identification and wholeheartedness. In *The Importance of What We Care about: Philosophical Essays*. Cambridge University Press.

Hieronymi, P. unpublished. Strawson's Descriptive Metaphysics of Morals and some upshots. https://ucla.app.box.com/v/StrawsonforScanlonfest. Last accessed: October 11, 2022.

Holroyd, J., R. Scaife, & T. Stafford. 2017. Responsibility for implicit bias. *Philosophy Compass 12*(3): e12410.

Nelkin, D. K. 2008. Responsibility and rational abilities: Defending an asymmetrical view. *Pacific Philosophical Quarterly 89*(4): 497–515.

Sifferd, K. L., & W. Hirstein. 2013. On the criminal culpability of successful and unsuccessful psychopaths. *Neuroethics 6*(1): 129–140.

Sripada, Chandra. 2016. Self-expression: A deep self theory of moral responsibility. *Philosophical Studies 173*(5): 1203–1232.

Strawson, P. F. 2008. *Freedom and Resentment and Other Essays*. Routledge.

Wolf, S. 1981. The importance of free will. *Mind 90*(359): 386–405.

11 Moral Responsibility, Free Will, and Determinism

In Chapter 10 we considered a few different theories about what internal, psychological causes should be seen as the ones that characterize responsible action. These theories are known as compatibilist theories, because they take moral responsibility to be compatible with determinism (in particular, by mechanistic psychological causation). There are some important worries that we have not yet considered about any compatibilist theory. To understand these problems we need to set compatibilism in the larger context of the debates about free will and determinism.

Free Will and Determinism

Causal determinism is the thesis that every event is necessitated by antecedent events and conditions together with the laws of nature. Or, to put it another way, the facts of the past, in conjunction with the laws of nature, entail every fact about the future. In a way, determinism *per se* is not the issue here – many philosophers on both sides of the debate think that quantum indeterminism means that "determinism" is unlikely to be true of our universe. Instead, it seems likely that quantum *indeterminism* is true in our world, which means that the actions of subatomic particles are not fixed by events of the past. Because of the strange nature of these particles, "[n]o superintelligence (not even God perhaps) could know the exact positions and momenta of all the particles of the

DOI: 10.4324/9781003402787-14

universe at a given moment because the particles do not have exact positions and momenta at the same time … hence their future behavior is not precisely predictable or determined" (Kane 2005: 9). But quantum indeterminism doesn't resolve the debate. Why not? Quantum indeterminism doesn't solve anything because the real worry is about whether we have *control* over the causes of our action and it's hard to see how the random behavior of subatomic particles could help with this. Whatever is true about electrons and protons, it's highly likely that our actions are caused by factors in the distant past ultimately beyond our control. Maybe those causal chains include some undetermined subatomic events, maybe they don't – either way, it seems like we're not in control.

Free will is not so easy to define and, indeed, people don't tend to agree about how to define it. Philosophers who think free will is compatible with determinism think about free will much differently from those who think the two are incompatible, as we will see.

We can divide the different theories depending on how they answer two questions: (1) If everything – including human action – were causally determined, would free will (in particular, the kind of free will needed for moral responsibility) be possible? And (2) Do we have free will? (see Table 11.1[1]). On the first question, incompatibilists think that if everything were determined, then there would be no free will. The second question divides incompatibilists into two groups: libertarians are incompatibilists who think that we do have free will (so, they don't accept determinism). Hard incompatibilists are incompatibilists who think we do not have free will (because either determinism is true and it is incompatible with free will, or there are uncaused, random events and that's also incompatible with free will). Compatibilists have a different answer to the first question: they think that determinism and free will are compatible. In answer to the second question, most compatibilists think we do have free will; they think that regardless of whether everything is causally determined, we might still have

Table 11.1 Two Questions about Free Will and Moral Responsibility

		If determinism is true, could we still have the kind of free will needed for moral responsibility?	
		YES	NO
Do human beings actually have the kind of free will needed for moral responsibility?	YES	Compatibilism (Hume, Strawson, Arpaly, Nahmias, Fischer & Ravizza, Nelkin, Wolf)	Libertarianism about free will (Kant, Ekstrom, Kane, O'Connor)
	NO		Hard incompatibilism/ Skepticism about moral responsibility (Schopenhauer, Caruso, Pereboom)

free will as long as our actions are caused in the right way. (We discussed various proposals for what the "right way" might be in the previous chapter – for instance, caused by one's real self or by one's reasons-responsive mechanism.) These are all viable theories, with contemporary defenders.

According to compatibilism, as we saw in Chapter 10, people are sometimes in a psychological state that warrants holding them responsible for what they do, even though that psychological state itself is caused. I have learned that however the definition is worded, it can be hard to get your head around what compatibilism really means. Many students I've talked to mistakenly think that compatibilism must mean putting together the idea that our actions have causal histories that extend beyond our control with the idea that we have the very special kind of uncaused free will that goes along with having an eternal soul. This is *not* what compatibilism says, which is a good thing since those ideas cannot be put together. Instead, compatibilism says that actions done freely, actions for which we are morally responsible, *just are* actions that are caused in a certain way (which way depends on the details of the compatibilist theory). If you are someone who is convinced it's impossible for free actions to be caused, you will find this jarring. It will seem like compatibilism is redefining free will and, actually, it might help if you think of it that way. In particular, it might help you to avoid confusion if every time you think about a compatibilist theory you replace "free will" with "the kind of free will worth wanting," as Daniel Dennett put it (1984), which means the kind of free will necessary for *moral responsibility* and the reactive attitudes like praise and blame.

You may ultimately decide to reject compatibilism, but don't make the mistake of thinking that it is incoherent or blatantly false without the need for any argument. In the previous chapter, we saw coherent proposals for what moral responsibility could mean even if determinism were true. Determinism is compatible with our actions being caused by the second-order desires that we wholeheartedly endorse, by our whole selves, or by our reasons-responsive mechanism or executive function. Determinism does not mean that these ordinary causes of responsible action are bypassed or hijacked by some external force.

Of course, even if compatibilism isn't incoherent, it could still be wrong. Historically, some philosophers have thought that it's terribly wrong. Kant called it a "wretched subterfuge" and the philosopher Wallace Matson said that the defense of compatibilism is "the most flabbergasting instance of the fallacy of changing the subject to be encountered anywhere in the complete history of sophistry" (cited by Fischer et al. 2007: 45). Intuitively, the obvious problem with compatibilism is that if our actions are caused by chains of events that go beyond our own choices, and we're not responsible for those chains of events, then we're not responsible for the actions that result from them. Why did Stella steal my car? Well, either she came from a family of thieves who didn't teach her that stealing was wrong, or some patterns of neurons fired because of the activity of previous neurons, which caused her to do it, or a combination of both of these things. In any case, why blame Stella? It's either her parents' fault, or the fault of physical events in her body that she has no control over! The problem here seems particularly troubling when it comes to punishment: why should Stella be punished if she didn't have ultimate control over what she did?

This intuitive problem is reinforced by some very forceful arguments, two of which we'll consider in the rest of this section. One very influential argument against compatibilism is the Manipulation Argument, which says that the problem with compatibilism is that it cannot distinguish between cases in which someone is manipulated into choosing to do something and cases in which someone chooses to do something because of deterministic causes ultimately beyond his control. If the compatibilist can't distinguish these two kinds of cases, then it looks like the determined person who chooses is no more free than the manipulated person, which is to say: not very free at all. Think again about the roommate who dances on your table (and breaks it) because she wants to and the roommate who breaks your table because she is hypnotized. The Manipulation Argument says basically this: The hypnotized roommate (Hyp) is not free, because she is determined to do what she does by the hypnotist, and therefore she is not morally responsible. If determinism is true, there is no relevant difference between Hyp and the un-hypnotized roommate (Unhyp): Unhyp is equally not free, because she is determined to do what she does by the facts of the past and the laws of nature. Therefore Unhyp is also not responsible for breaking your table. Since the point generalizes, if determinism is true, no one is ever responsible for what they do, and compatibilism is false.

We have already seen the resources that compatibilists have for answering this kind of challenge. The philosophers we discussed in Chapter 10 reject the premise that there's no relevant difference between Hyp and Unhyp. They think they have explained the relevant differences with their theories: Unhyp has the relevant second-order, wholeheartedly endorsed desire, while Hyp does not, or Unhyp is responsive to reasons and Hyp is not, or there is some other psychological difference between the two. You may think that compatibilists have not answered this objection yet, because you think they haven't got the psychological conditions quite right. In this case, you would be a compatibilist who thinks there is more work to be done on the theory – and you would be in very good company! On the other hand, you may think that the Manipulation Argument could be pushed further so that even with these complex facts about our psychologies in hand, there is still no meaningful distinction between someone who is manipulated to choose and someone who chooses for her own reasons but is ultimately determined to do so.

One way to push the Manipulation Argument further comes from Al Mele's (2006) discussion of Ernie.[2] Here's the basic story: decades ago, the goddess Diana created a zygote, which developed into Ernie who is now a person with normal executive function and reasoning capacities. When Diana created the zygote, she used her knowledge of the laws of nature and all the facts about the world to ensure that this zygote, decades later, would steal an apple from Target. Now in his thirties, Ernie steals an apple from Target. Given that the chain of events that leads to Ernie's action crucially includes the agency of the goddess, Ernie does not seem responsible. But now, the objection continues, people who were created in the usual way, not by goddesses, are just the same as Ernie because we also have been determined to do what we do by a chain of events that is beyond our control.

Here is how Mele summarizes the challenge to compatibilism:

1. Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
2. Concerning free action and moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
3. So determinism precludes free action and moral responsibility (Mele 2006: Kindle location 2771).

The Zygote Argument causes real problems for compatibilism. Nevertheless, compatibilists have replies. Kadri Vihvelin (2022) suggests that we might reject the first premise and decide that, actually, Ernie *is* morally responsible for what he does. Oisin Deery and Eddy Nahmias (2017) take issue with premise 2 and argue that there is a significant difference between determinism and manipulation (say, by goddesses or brain surgeons) when it comes to moral responsibility. Briefly, they say that the two cases are different in terms of which cause makes the *biggest* difference to the action: in the case of Ernie, it's Diana's actions that are the most important part of the explanation; whereas, for regular folk who developed from what Mele calls "normal human zygotes," it's our desires and choices that make the biggest difference.

Now, you may still be suspicious about compatibilism. You may be thinking that spelling out the psychological differences between manipulated agents and regular agents is beside the point, because as long as we're talking about psychological *causes*, something is missing. To see what might be missing, let's turn to a second type of argument against compatibilism. According to the Alternative Possibilities or Forking Paths Argument, the problem with compatibilism is that free will (and hence moral responsibility) requires the existence of genuine alternatives (or paths branching off from the present) from which we can choose, but determinism precludes genuine alternatives. Unhyp is only free not to dance on and break your table if at the moment when she chose to dance on the table she could have stayed on the floor. She is, therefore, only responsible for breaking your table if at the moment she chose to dance she was at a fork in the road from which she could choose to go either way.

Now, the first thing to notice about this argument is that compatibilists have a response. Recall that this question about forking paths is just what is at issue in the distinction between regulative control and guidance control, discussed in the previous chapter. The former (regulative control) requires genuine alternatives, the latter (guidance control) does not, and compatibilists like Fischer think that only the latter is necessary for moral responsibility.

Harry Frankfurt, also discussed in the previous chapter, agrees with Fischer. His argument for thinking that we do not need genuine forking paths for responsibility relies on cases that have come to be called "Frankfurt-style cases" (Frankfurt 1969; see also Fischer 1982). A Frankfurt-style case is a case in which a person – say, Stella the car thief – does what she does from her own motives and for her own reasons, *but* also has a mechanism in place, implanted by a nefarious brain surgeon, whose function is to guarantee that Stella would steal my car even if she chose not to. So, the idea is that Stella would have stolen my car even if she didn't want to, but in fact she did want to and did the deed *because* she wanted to! She could not have done otherwise, and yet it seems like she is responsible because she is acting on her own desires. Compare Stella to Ella, who had the same surgery as Stella but who does not want to steal my car. Ella considers her reasons and decides not to steal my car, *but does it anyway*, because of the mechanism implanted by the nefarious brain surgeon. Are they equally responsible? If you think Stella is more responsible than Ella, then you're on the side of compatibilism: genuine forking path alternatives are not needed for responsibility.

The crucial question now is this: do free will and responsibility require forking paths or genuine alternatives? What kind of question is this? And how do we decide? We can't discover whether free will requires forking paths by empirical investigation. We might be able to decide whether we really do have genuine alternatives by appealing to empirical science (more on that later), but the question we're asking here is a conceptual one, not an empirical one. Is the best understanding of free will one that requires genuine alternative forking paths? How do we answer a question like this? Recall the methodology of reflective equilibrium, according to which what we are doing when we ask deep philosophical questions like this is reaching a conclusion that puts our intuitions, principles, and theories into a coherent whole. If this is what we have to do to answer the question, it seems that we need to know more about what our relevant intuitions actually are.

Intuitions and Experimental Philosophy

Our question is this: Is compatibilism supported in reflective equilibrium, or is compatibilism too much at odds with our intuitions to be the right theory? We can start by asking what *your* intuitions are. Consider the following case of Fred and Barney:

> Imagine there is a world where the beliefs and values of every person are caused completely by the combination of one's genes and one's environment. For instance, one day in this world, two identical twins, named Fred and Barney, are born to a mother who puts them up for adoption. Fred is adopted by the Jerksons and Barney is adopted by the Kindersons. In Fred's case, his genes and his upbringing by the selfish Jerkson family have caused him to value money above all else and to believe it is OK to acquire

> money however you can. In Barney's case, his (identical) genes and his upbringing by the kindly Kinderson family have caused him to value honesty above all else and to believe one should always respect others' property. Both Fred and Barney are intelligent individuals who are capable of deliberating about what they do.
>
> One day Fred and Barney each happen to find a wallet containing $1,000 and the identification of the owner (neither man knows the owner). Each man is sure there is nobody else around. After deliberation, Fred Jerkson, because of his beliefs and values, keeps the money. After deliberation, Barney Kinderson, because of his beliefs and values, returns the wallet to its owner.
>
> Given that, in this world, one's genes and environment completely cause one's beliefs and values, it is true that if Fred had been adopted by the Kindersons, he would have had the beliefs and values that would have caused him to return the wallet; and if Barney had been adopted by the Jerksons, he would have had the beliefs and values that would have caused him to keep the wallet.
>
> (Nahmias et al. 2007: 38–39)

Do you think Fred and Barney acted of their own free will? Do you think Fred is morally responsible – deserves to be blamed – for keeping the money that isn't his? Do you think Barney is morally responsible – deserves to be praised – for returning the money? At this point your intuitions might be influenced by what you have already read about compatibilism and incomptabilism. Perhaps some of you were convinced by the theories considered in Chapter 10 and are wondering what Fred and Barney's second-order desires are, or whether they are responding to reasons sufficiently in this particular case. Perhaps some of you, having been convinced by the arguments against compatibilism presented at the beginning of this chapter are thinking that there's no way Fred and Barney could be responsible since they could not really have done otherwise than what they did.

Historically, philosophers have used their own intuitions as the inputs to the reflective equilibrium method. This isn't a bad thing to do – after all, philosophers have thought hard about the problems and have the resources to make relevant distinctions. However, philosophers' intuitions don't all agree: some incompatibilists say that compatibilism is the most counterintuitive theory on earth; compatibilists think it's perfectly fine. Furthermore, when it comes to the question of what free will is and what moral responsibility requires, there is an additional problem with relying on philosophers' intuitions alone. What (most) philosophers are trying to understand when they try to understand the kind of free will that is necessary for moral responsibility is something that all people (not just philosophers) would recognize as the thing that we worry about when we wonder if someone should be punished or when we read articles in the newspaper that tell us neuroscientists have proved there is no free will. Philosophers want to be talking about the same thing that everyone else is talking about! It doesn't add much to the defense of a theory of moral responsibility to say that it is intuitive to the philosopher who constructed it.

Given this, some philosophers have started to take reflective equilibrium to the streets (or the lab). Instead of relying on their own intuitions about free will and responsibility, they find out what "normal" people think by conducting studies that ask people about their intuitions. This work is in the relatively new field of philosophy called "Experimental Philosophy." In fact, the case of Fred and Barney is taken from an experimental philosophy article by Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner (2007). What they found when they surveyed people who had not studied the free will debate is that significantly more people agree than disagree that Fred and Barney acted freely and are responsible for what they did.[3] This research makes it look like compatibilism isn't so unintuitive after all.

But not all of the experimental philosophy research favors this conclusion. Shaun Nichols and Joshua Knobe (2007) argue that people have compatibilist intuitions when their intuitions are distorted by emotional responses, but that when people think more carefully and deliberately about the question, they have incompatibilist intuitions. Nichols and Knobe make this argument using the same methods as Nahmias, but with different scenarios. They give all their participants descriptions of a causally determined universe called "Universe A" and then they divide the participants into two groups. The "Concrete" group gets this question:

> In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and three children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.
>
> Is Bill fully morally responsible for killing his wife and children?

The "Abstract" group gets this question:

> In Universe A, is it possible for a person to be fully morally responsible for their actions?

What they found was that in the "Concrete" group, 72 percent of the people said that Bill is responsible (this is the compatibilist answer, since Bill's actions are determined). In the "Abstract" group, 86 percent of people said no, it is not possible for a person to be fully responsible in a deterministic universe (the incompatibilist answer).

Nichols and Knobe hypothesize that what's going on here is that those who read the scenario about Bill are swayed by their emotional response to Bill's bad actions, while those who read the abstract question think more abstractly and hence more clearly about the compatibility of moral responsibility and determinism. Given what we learned in Chapter 6 about the connection between moral judgments and emotions, one way to respond to Nichols and Knobe would be to point out that emotions do not necessarily mislead us. Emotional responses are integral to judgments of moral responsibility, one might say, so we do not think more clearly without our emotions in this domain. Think back to Strawson and Wallace: attributions of moral responsibility are normative judgments. If normative judgments essentially involve emotion, then it isn't surprising that emotion changes our intuitions. In this way of thinking, our emotions aren't leading us to make mistakes. Nichols and Knobe do consider this possibility, and they have designed other studies to try to sort out which hypothesis is better. Interested readers should look to the suggested readings at the end of this chapter and follow the progress of this line of research by searching for the latest.

Nahmias has a different objection to Nichols and Knobe. He objects to the way that they describe the deterministic universe, Universe A. Nahmias thinks that when they say that in Universe A "given the past, each decision *has to happen* the way that it does" and contrast that with Universe B in which "each human decision *does **not** have to happen* the way that it does," this leads many people to think that agents in Universe A will do what they do, no matter what they want to do or what they decide to do. But this is not the case: in a deterministic universe, our desires are part of the causal chains that lead to actions. There is a difference, then, between determinism (everything is caused) and *bypassing* (everything is caused by a series of events that bypasses our psychological states such as desires and choices). Once we understand the difference between the claim that all actions are causally determined and the claim that actions are determined in such a way as to bypass our psychological capacities, we can see that bypassing is the biggest threat to our ideas of free will and moral responsibility. Bypassing in one form is the problem we considered in Chapter 10 when we discussed the challenge from psychology to the idea that we are reasons-responsive creatures. There we concluded that while we might be less rationally capable than we once thought, we are not completely hopeless either. If the real problem is bypassing, compatibilism remains a viable position – because even if determinism is true, it is *not* true that our psychological processes are bypassed when we act. And it is plausible to think that bypassing (not determinism) is the real problem for attributions of responsibility. After all, the hypnotized person seems to have her beliefs, desires, virtues and so forth bypassed (by the actions of the hypnotizer), and that's why we do not think she is free or responsible for what she does under hypnosis.

Nichols again, this time with collaborator David Rose, argues that bypassing isn't actually driving the bus when people make judgments about moral responsibility. Instead, they argue (using more experimental philosophy) that it's the incompatibilist intuitions that make people think psychological processes were bypassed in the first place. People don't think choices and decisions could have happened in a universe in which everything is determined, they say; the intuitive roots of incompatibilism are very deep (Rose & Nichols 2013).

Other philosophers have argued that there are just many different intuitions about free will and determinism. For example, Ivar Hannikainen and colleagues (2019) show, in a large study of over 5,000 participants in 20 countries, that intuitive judgments about moral responsibility and determinism are quite variable; there may be no "predominant intuition" on this topic. The debate about our intuitions about moral responsibility and free will is ongoing. There is even now a whole anthology dedicated to advances in experimental philosophy on the topic of free will (Nadelhoffer & Monroe 2022).[4]

It's difficult to draw any definite conclusions from experimental philosophy at this point. We do know that there is some evidence for thinking that people accept compatibilism as long as some of our psychological capacities are part of what causes us to decide what to do. There is also evidence for thinking that in some circumstances, people are disturbed by determinism. We have not reached an obvious conclusion. Some things might become more clear as research progresses, but it seems unlikely that one answer will turn out to be a slam dunk. Where do we stand? There are two important points to consider here.

First, we can conclude that someone who wants to argue for incompatibilism should not begin by assuming compatibilism is counterintuitive without addressing the current research in experimental philosophy. It at least cannot be taken as obvious that compatibilism about determinism and moral responsibility is a "wretched subterfuge." Second, we must remember that intuitions alone do not decide what free will and moral responsibility really are. Intuitions about cases are relevant to our theories because as philosophers we want to make sure that our theories track something meaningful to ordinary people. However, there are other factors that go into justifying a theory of moral responsibility. One of these other factors is the coherence of our theory of moral responsibility with scientific theories, some of which we'll consider later in this chapter. Another factor is the cost of abandoning the idea of moral responsibility. This is the point Strawson made about how important the reactive attitudes are to us and to how we relate to each other. Of course, compatibilism isn't the only option that allows us to retain our notion of moral responsibility; one might take the view that we are sometimes morally responsible, even though responsibility and determinism are incompatible. This view is called libertarianism, a version of incompatibilism, the topic of the next section.

Two Kinds of Incompatibilism

Compatibilists make sense of moral responsibility by identifying responsible action with action that is caused by certain psychological features of the person, but there is another way that we have not yet considered. One might reject determinism and argue for a non-deterministic or "libertarian" conception of free will and moral responsibility according to which we have an undetermined power to choose. In its earlier incarnations, libertarianism made free will rather mysterious. The idea that agents have special powers that stand outside of the causal order of the universe – sometimes called "agent causation" – is hard to believe for scientifically minded people.[5] Moreover, libertarians had to answer Hume's challenge: they had to explain why we would be held responsible for actions that are not caused by our character or determined by our beliefs and desires. In other words, if we have a will that is itself truly uncaused, how do *we* get any blame or credit for it? Why isn't it just some weird random force doing what it does arbitrarily? If the answer is, *we* determine what our free will does, who is the *we*, if it isn't our beliefs, desires, and psychological processes?

It is because of these problems (together with problems for the compatibilist option) that some philosophers, such as Derk Pereboom, have opted for "hard incompatibilism": the view that whether or not determinism is true, we do not have free will. We do not have it if determinism is true, because free will and determinism are incompatible (hard incompatibilists believe that the Manipulation and Forking Paths arguments are successful against even sophisticated versions of compatibilism). And we do not have it if determinism is false, because indeterminism at the quantum level doesn't help and because special, undetermined agent causation is incompatible with our best scientific theories (Fischer et al. 2007: 85). Hard incompatibilists also deny that we have moral responsibility in the sense that supports backwards-looking judgments of praise, blame or deservingness. For this reason, hard incompatibilists – also called "moral responsibility skeptics" – reject backwards-looking or *retributivist* justifications for punishment. Retributivism is the view that punishment is justified by the fact that a person who performs certain kinds of actions (such as actions that harm other innocent people) *deserve* to be punished. Retributivism contrasts with forward-looking justifications of punishment, which consider the *consequences* of punishment, such as deterrence or rehabilitation.

You would not be alone if you worried about giving up moral responsibility, deservingness, and retributive justice. Critics worry that believing hard incompatiblism would seriously undermine many desirable aspects of our morality. Would it? In his defense of hard incompatibilism, Pereboom outlines some of the good consequences of giving up the idea that people are fundamentally responsible for what they do (Pereboom 2001; Fischer et al. 2007). For instance, there will be less moral anger, and anger is a painful emotion that often does more harm than good. Neil Levy (2015), another moral responsibility skeptic, even argues that our system of punishment would be improved by giving up on desert-based moral responsibility, basically because we could achieve better results for society with a system that punished people less severely than is required by retributivism. Furthermore, Pereboom argues, not as much would be lost by giving up our ordinary notion of moral responsibility as someone like Strawson suggests. Many of our emotional responses and ordinary ways of relating to other people would be unscathed if we did not believe people are praiseworthy for their good deeds or blameworthy for their bad ones. We would still feel joyful when good things happen, sad when bad things happen, and love for our friends and family.

Experimental philosophers have jumped into this debate, too. In order to investigate what would happen if people believed hard incompatibilism, they prime people to diminish or augment their belief in free will. There is some evidence that when people are encouraged to believe that there is no free will (hard incompatiblism), they become more aggressive (Baumeister, Masicampo, & DeWall 2009), and more biased against an out-group (Zhao et al. 2014). But other studies show some positive effects of skepticism about free will and responsibility, such as less vindictiveness (though, interestingly, only among women! See Caspar et al. 2017), authoritarianism and punitiveness (Carey & Paulhus 2013). Chances are that the impact of belief or skepticism about moral responsibility on our lives will be a mixed bag, but we certainly need more evidence here.

If you aren't keen on hard incompatibilism, but you don't like compatibilism either, you're in luck, because moral responsibility skepticism is not the only viable incompatibilist option. New and improved forms of libertarianism have done better than their predecessors.[6] These theories take free actions to be *non-deterministically caused* by some event in the agent like a decision, and so they are sometimes called "event-causal theories." The terminology can be confusing if "causal" makes you think of determinism. Libertarian theories of free will reject determinism, but they do not reject the idea that our actions are caused. This is how they find the balance between actions that seem beyond our control because they are deterministically caused and actions that seem beyond our control because they are the random result of chance.

The idea of non-deterministic causation is important here. Laura Ekstrom, who defends an event-causal theory of free will, characterizes it this way:

> The crucial matter concerning indeterministic causation, whichever theory best captures its nature, is that, if there is such a thing, then some events are causally related to their effects without necessitating them. Events that indeterministically cause other events make a difference for those effects, but the effects might not have occurred, in the same circumstances and holding fixed the natural laws.
>
> (2019: 131)

Defenders of event-causal theories of free will make use of non-deterministic causation to capture the essence of free will. Ekstrom (2019) argues that free actions are caused – non-deterministically – by the agent's reasons:

> A decision or other act is directly free just in case it is caused non-deviantly and indeterministically by reasons of the agent's – such as convictions, desires, values, beliefs, and preferences – and other reasonable compatibilist conditions on free action are met, including that the act is not compelled and is not the result of (non-self-arranged) manipulation or coercion.
>
> (137)

Crucially, according to Ekstrom, because the agent's reasons cause her action without determining it, there was – at the moment she chose – space for her to do other than she did. In other words, what Ekstrom is saying is that when someone decides to do something, it isn't random. Rather, her decision is caused by her reasons in the sense that her reasons increase the probability that she will decide in a way that those reasons support.

Robert Kane defends a similar version of libertarianism. Kane's theory makes use of the science of chaos theory and quantum indeterminism in order to argue that there is some "space" in the causal network for genuinely free actions. Kane does not think free choice about our actions requires any special power. Rather, like Ekstrom, he thinks we choose how to act in the way that compatibilists think we do: by reflecting on our reasons and endorsing one path over another. But Kane and Ekstrom (unlike the compatibilists) thinks there are genuine forking paths and that we (at least sometimes) determine what we do *by deciding* which path to take.

In Kane's view, when we make certain kinds of choices we determine who we are by engaging in reasoning and then endorsing one option over the others. Such choices are not arbitrary, because we make them for reasons; nevertheless, the reasons do not determine which choice we will make. Writing in the voice of someone who has the kind of free will he is describing, Kane says

> I did have *good* reasons for choosing as I did, which I'm willing to stand by and *take responsibility for*. If these reasons were not sufficient or conclusive reasons [they were not reasons that determined my choice], that's because, like the heroine of the novel, I was not a fully formed person before I chose (and still am not, for that matter). Like the author of the novel, I am in the process of writing an unfinished story and forming an unfinished character who, in my case, is myself.
>
> (Kane 2007: 42)

In this way of thinking, when a person with free will finds a wallet, he or she decides whether to be a sinner or a saint. There are reasons on both sides and, whichever the person chooses, he or she will have chosen freely if and only if, at that moment, it is not causally determined what that person will do. Now, not every single action we perform is like this. Kane does think that much of what we do is determined by our character and does not involve the exercise of our free will. But we are ultimately responsible for the actions that result from our character only because we create our character through freely chosen self-forming actions (such as, for example, returning the wallet).

Now, one big question for this variety of libertarianism is whether there really is the right kind of indeterminism, the kind that makes room for us to exercise free will by endorsing one set of reasons or another. Kane describes the condition as "a kind of 'stirring up of chaos' in the brain that makes it sensitive to micro-indeterminacies at the neuronal level" (Kane 2007: 26). Whether this kind of indeterminacy exists is an interesting and important question, but it is an open empirical question beyond the purview of this book. (Interested readers are encouraged to read more of Ekstrom's and Kane's work, cited in the bibliography). Another big question for event-causal libertarianism is how indeterminism – even if it occurs in the right place – solves the intuitive problem we started with, namely, the problem that we don't have control over the causes of our actions. If my reasons may or may not indeterministically cause me to decide to steal your car, how am I ultimately in control? How does non-deterministic causation or self-forming choice escape deterministic causation by my psychological capacities without leaving my control altogether and without embracing the somewhat mysterious agent causation?

Interestingly, because Kane's and Ekstrom's versions of libertarianism also make use of ideas about our ordinary rational capacities, they face some of the same questions that compatibilists have to answer: which rational capacities are the relevant ones, and how often do these capacities actually play the right causal role in our choices? We have already considered (in Chapter 10) some psychological evidence that we use our rational capacities less frequently than we think. This evidence is not really a problem for libertarians either, since they don't think we exercise free will all the time. But there is a different scientific challenge to the idea that we use our rational capacities to make choices, that targets libertarianism as well as compatibilism. We'll consider this challenge in the next section.

The Challenge from Neuroscience

Some neuroscientists have claimed to show that there is no free will by investigating the brain. The claim is that when you look at what is going on in the brain, you discover that our decision-making capacities are *not* part of what causes us to choose what we choose; the brain processes that cause action are distinct from whatever in the brain corresponds to consciously making a choice. If this is correct, libertarianism would be in trouble because there would be evidence that brain processes determine what people will decide before they are aware of making a choice – that is, before our reasons could (indeterministically) cause us to act. Compatibilism would also be in trouble, of course, because it assumes that free and responsible actions are (deterministically) caused by those decision-making capacities.

The pioneering studies in this area were conducted by the psychologist Benjamin Libet. Libet and his colleagues (1985) would have participants sit at a desk in front of a dial with a quickly rotating arm (like a fast moving second hand on a clock). Participants were asked to flex their wrists at some point, whenever they chose, and to note the precise location of the arm on the dial when they were first aware of a wish to flex. These participants had electrodes on their heads that were attached to an electroencephalogram (or EEG) that measures electrical activity at the surface of the brain called "readiness potentials" (a proxy for neuronal activity). Libet discovered that the readiness potential (a sign that a person is ready to act) that preceded the action of wrist flexing happened *before* the person was aware of any desire or decision to flex (according to the participant's report about the position of the arm on the dial). Subsequent research has confirmed and extended Libet's findings. For example, John Dylan Haynes and colleagues replicated Libet's results using fMRI methods instead of EEG technology (Soon et al. 2008). The conclusion of all this research, according to one researcher quoted in *WIRED*, is that "Your decisions are strongly prepared by brain activity. By the time consciousness kicks in, most of the work has already been done" (Keim 2012).

This research is extremely interesting, but philosophers have not been impressed with it as an attack on free will. Indeed, it seems like one of the rare points of agreement among philosophers that neuroscience does not disprove the existence of free will. Philosophers have argued that these neuroscientific studies do not actually show that our decisions play no causal role in determining how we act. For example, Adina Roskies, a philosopher and neuroscientist, argues that Libet's experiments do not correctly measure the time that we will something to happen. Rather, she says, they measure "the meta-state that is the awareness of one's own conscious intention" (2021: 168); in other words, by measuring when people say they recall willing an action, Libet is tracking a different state from the actual willing, which may come earlier. Al Mele (2006) locates the problem with the experiments in a different place. He argues that the right interpretation of the data is that the brain activity that precedes conscious decision is likely just a preconscious urge to flex, not an intention to do so. These preconscious urges are sometimes acted on and sometimes overridden by conscious decision making. This is consistent with Libet's findings, according to which readiness potentials may be vetoed by conscious intention (which is why Libet admits that we do have "free won't"). If Roskies or Mele is correct (or if they both are), then scientists have not proven that there is no free will.[7]

Furthermore, compatibilists and event-causal libertarians like Ekstrom and Kane assume that our rational capacities shape what we do overall, but they do not require that we employ our rational capacities every time we do something. For example, as I am writing this chapter, my actions are governed by many of my conscious intentions and decisions of varying degrees of specificity (to write a book, to write a chapter on free will, to consider the evidence from Libet, to sit at my desk in front of my computer and so on), but many of the things I'm currently doing are not things I notice any intention to do. I just hit the "period" key, for instance, but I didn't make a decision to do that. Nevertheless, my hitting that key (and the "y" key just now!) is – in the big picture of what I'm doing today – governed ultimately by my decision to write this book. Similarly, what we might say about Libet's participants is that even if their wrist flexings were *not* consciously chosen, their actions are (in the big picture) governed by their decisions to be in his experiments, to follow Libet's instructions and so on (Nahmias 2012b).

It would be a problem for free will and moral responsibility if, when we looked at the big picture, we found that our conscious intentions are always just momentary blips that enter the causal stream after other bits of our brains (say, readiness potentials) have already set action in motion. For compatibilists, this would not leave room for the right kind of causal history where our rational capacities are a more integrated part of the causes of our actions. For event-causal libertarians, this would not leave room for our reasons to sometimes indeterministically cause our actions.

The compatibilist Daniel Dennett has a helpful way of explaining why this "blip" pictures isn't the right one. According to Dennett (2004), there can't be one place in the brain that is the place where conscious intentions to act for reasons are located: the conscious intentions we look for when we attribute responsibility are not little blips in the brain. Representing reasons to yourself, considering these reasons, forming an intention to do something on the basis of one of those reasons, and remembering all of this when you are doing what you intended to do – these are highly complex and varied activities that require many different parts of the brain. The self who acts for reasons is spread out across the brain, and so too are responsibility and free will. It is a mistake, according to Dennett, to think that *you* (the responsible self) must be at some specific point deciding what to do with some event that happens in your brain. And if it doesn't make sense to think of conscious decisions as very localized events in the brain, then it doesn't make sense to argue that conscious decisions happen after actions are set in motion.

What we have learned from neuroscience so far, then, does not render irrelevant the big questions that both compatibilists and libertarians have to answer. These are questions about which rational capacities are relevant and what counts as using them sufficiently well for the person to count as responsible. Indeed, these questions seem to have increasing importance now that science is discovering more about normal and abnormal brains, as we will see in the final section of this chapter.

Can I Be Excused?

A good theory of moral responsibility will tell us when people are responsible and also when they ought to be excused for what they've done. We have already seen some examples of people who should be excused from responsibility and blame, such as the two-year-old hitter and the hypnotized table-dancing roommate (Hyp). These cases are fairly easy, because it is clear that in *any* theory of the relevant rational capacities the two-year-old doesn't have them yet, and Hyp's seem to have been completely bypassed. But there are many more difficult cases.

In the early 2000s, a schoolteacher (he wasn't named in reports, let's just call him Teacher) suddenly developed tendencies toward pedophilia. He began visiting child pornography websites, molested his stepdaughter and expressed fear that he would rape his landlady. Shortly after these behaviors began, Teacher checked himself into the hospital complaining of terrible headaches. It turned out that he had an egg-sized tumor in his brain and, after this tumor was removed, his perverse sexual tendencies disappeared. Was Teacher responsible for what he did while he had the tumor?

In 1983, the serial killer Brian Dugan kidnapped, raped, and killed a ten-year-old girl. He had already committed several murders, but for this particularly heinous crime the prosecuting attorneys sought the death penalty. Dugan's lawyers argued that he had a mental illness – psychopathy – that excused him from responsibility for his crime. Dugan did indeed fit the profile of a psychopath: he scored in the 99.5th percentile on the psychopathy checklist and brain scans showed that his brain was abnormal in the way that many psychopaths' brains are abnormal. Assuming that Dugan is indeed a psychopath, is he responsible for what he did?

Both Teacher and Dugan have abnormal brains, though Teacher's brain was only temporarily abnormal, while Dugan's brain is congenitally different. As brain science develops, we will know more and more about the brains of people who commit crimes and the attempt to use information about people's brains as a legal defense is becoming well known. If you're interested, you can search "brain scans as defense" or "my brain made me do it" on Google and find many articles and blogposts on this topic. There is even a name for the field of study that investigates these issues: neurolaw. We cannot delve too deeply into the legal questions here, but two points about the legal context are worth mentioning before we return to the issue of moral responsibility.

First, it turns out that people are very influenced by brain science. Adina Roskies (2008) has argued that we have reason to be cautious about the use of brain scan images in public discourse, because people tend to treat pictures as providing a very direct kind of evidence that brain scan images do not, in fact, provide. Unless there is some explicit reason to suspect tampering, we tend to think of pictures as on a par with eyewitness testimony. Video surveillance footage, for example, is considered to be excellent evidence (though in this era of deep fakes we have to be more cautious than before). Because we think of pictures this way, we tend to think that pictures of brain scans provide similarly convincing evidence. But, as Roskies argues, there is a lot of distance between the brain scan image and any conclusion that might be drawn from it, and this distance must be bridged with inferences made by experts who are drawing on their own imperfect knowledge and theories of the brain. Brain scan images are not like video surveillance footage, but people tend to treat them as the same. Therefore, this evidence introduces a risk that people will put more weight on it than is warranted. This is something that should be remembered by anyone interested in neurolaw. If neuroscientific evidence biases us, it must be handled with a great deal of care.

Second, the question of whether Teacher and Dugan are legally responsible is different from the questions of whether they are morally responsible. If we decide that they are not morally responsible, it does not follow that they should be let go on the streets. Deterrence is a perfectly good reason for detaining people in prisons or institutions, and it does not necessarily depend on establishing moral responsibility in a sense that supports retributive punishment. On deterrence grounds, we are justified in depriving someone of their liberty if they are a serious threat to the rest of us. (Of course, the criminal you punish should be causally responsible – it wouldn't do to punish someone for something he didn't do in any sense at all.) Indeed, those who think that we must abandon our ordinary understanding of free will and moral responsibility – the hard incompatibilists, mentioned above, who think we are never morally responsible – believe that a just set of laws will include a system of punishment focused on deterrence.

Now we can return to the question about moral responsibility. When should people be excused from blame? Does having an abnormal brain change whether you are responsible? Does it matter whether the abnormality is a temporary affliction or a permanent condition? Do the theories we have considered give us any help with these difficult questions? And what about conditions of the brain that are not currently recognized as abnormal or diseased? What if Bernie Madoff's lawyer could show that his brain exhibits patterns typical of swindlers and that therefore he could not have helped cheating people out of their money? His brain made him do it!

The first thing we should do is to stop talking about our brains making us do things. If you believe, as many philosophers and scientists do, that the mind is the same thing as the brain, then two things are true. First, your brain will always be involved in whatever you do, but it doesn't *make* you do anything; rather, your brain exhibiting certain kinds of neuronal patterns *just is* you doing something. It's not like your brain is an independent force that could rebel against you and run away with your body. Second, when people act well or badly, these differences will appear as differences in their brains. So, if Bernie Madoff chose to swindle old people out of their money, this will show up in his brain (we may not have the tools to observe it, but his desires and beliefs will in fact be there, in his brain). Far from being an excuse, if Bernie Madoff could show that his desire to cheat people out of their money is there in his brain, he would just be providing evidence that he really is a bad guy.

Now we can return to our question about the conditions for excusing people from blame. Consider the case of psychopathy first. Self-expression theories seem to reach different conclusions about psychopaths than reasons-responsiveness theories. Consider the psychopath whose real self is psychopathic.[8] His second-order desires endorse his first-order immoral desires and he has no intervening wholehearted desires to be a better person or to be nicer. His affective and cognitive capacities are not in conflict; the psychopath has an integrated bad will. So, if these claims about the psychopath are correct, according to the views that define moral responsibility in terms of a conception of the self, psychopaths are responsible for what they do. (This may over-simplify matters; it's open to self-expression theories to include other conditions on responsibility that render at least some psychopaths not responsible for their actions).

Reasons-responsiveness theories (and probably also libertarian theories like Ekstrom's and Kane's, given their emphasis on reasons) will not necessarily draw this conclusion, because there is evidence that some psychopaths lack certain rational capacities (Kennett 2006). If the psychopath in question does not have any capacity to conform his actions to what are good reasons, or if he is not able to grasp certain reasons, including moral reasons, at all, then reasons-responsiveness compatibilists may conclude that this psychopath is not morally responsible for what he does, depending on which capacities the reasons-responsiveness theorist thinks are important. Moreover, it seems clear that not all psychopaths have the same deficiencies and this will matter to their moral responsibility. For example, Sifferd and Hirstein (2013) provide evidence that psychopaths differ in what capacities they have for executive function, which include the capacities to grasp and act for reasons.[9] Psychopaths who lack executive function – unsuccessful psychopaths, as they call them – should not be held responsible for their actions; however, "successful psychopaths may be fully culpable, because they possess the executive functions to allow them to notice and correct for their criminal tendencies" (2013: 130).

When it comes to temporary brain abnormalities, such as Teacher's tumor, the self-expression theories suggest a different conclusion. What we might think is going on in Teacher's case is that his tumor produces first-order desires that conflict with what he wholeheartedly endorses. He wants to look at child pornography because of the tumor, but he really does not want himself to have this desire. If this is the case, the theory would say that Teacher's actions should be excused insofar as they are caused by the tumor rather than by him. Reasons-responsiveness theories suggest the same answer insofar as Teacher's ability to grasp reasons is impaired by the tumor or insofar as his capacity to act for reasons is entirely bypassed by these alien desires. As for the whole self theory, if Teacher has a new desire, then it is very likely to be marginalized in the rest of his psychology and so he has diminished but not zero responsibility for acting on it.

These cases raise many more questions. If psychopaths are wantons (who have no second-order desires or none that can influence their actions), then the self-expression theory would say they are not responsible after all. Do the particular impairments that psychopaths have render them incapable of having second-order desires or wills? If rational capacities come in degrees, how much executive function must a psychopath have to be considered a "successful psychopath"? Indeed, what degree of rational capacity must a developing non-psychopath have to be considered morally responsible? Four-year-olds aren't morally responsible, but what about twelve-year-olds or sixteen-year-olds? Where is the threshold for moral responsibility and is it the same across contexts? We may want to hold a ten-year-old responsible in the sense that warrants a time-out, but not in a sense that warrants incarceration.

Different complications arise when we think about conditions that are temporary or changeable. Consider the possibility of an anti-psychopathy medication. If psychopathy could be treated, would we say that the person on the drug is "more himself" (and hence more responsible, on the self-expression view) than the person off the drug? If we could cure poor executive function with drugs, would we require people to take this medication and punish them if they don't? Pharmaceuticals introduce their own complications. For example, consider the case of Hannah (not her real name), a psychology professor whose behavior became erratic and dangerous after she was prescribed a Parkinson's drug for her depression. Bioethicist Carl Elliott describes her case:

> The longer Hannah took pramipexole, the worse her behavior became. She picked fights in bars. She bought provocative clothes at a shop that sold outfits to streetwalkers. She wore a T-shirt that said "Fuck Cornell" and bought a gold necklace that spelled out B-I-T-C-H. She insulted friends and family members. She took a box cutter to several paintings by a close friend. She euthanized her cat on a whim … Her driving became dangerously reckless. "I passed a double-decker bus going 80 miles an hour around a blind turn" …
>
> (Elliott 2022)

Once weaned off the drug, Hannah and other patients taking it were mortified by their past behavior. Hannah called it "total degradation." Is Hannah responsible for what she did while on the drug? Should she apologize to the friend whose paintings she ruined? She feels ashamed of what *she* did – should she?

These are difficult questions and the answers will depend on the details of the case. The important thing for our purposes is to notice that a theory of responsibility will direct your investigation into these details, telling you what questions to ask and which facts to pay attention to. If you think the self-expression theory is correct, you will want to know whether Hannah is capable of forming preferences about her preferences: Can she form a conception of the kind of person she wants to be that is effective in her actions? If you think the reasons-responsiveness theory is correct, you will want to know about the Hannah's rational capacities: is she able to recognize patterns of reasons and use these to decide what to do? If you think Susan Wolf is correct, you will want to ask whether the person has developed in a way that makes her rational capacities track the truth about the world, including truths about morally good and bad ways of treating others. If you think Kane's libertarianism is right, you will want to know whether the person's action was the result of the character she created through free, self-forming actions (or whether the action was itself an undetermined choice to act for certain reasons rather than others). If you think hard incompatibilism is the right theory, you will need to ask what it is about different people that justifies treating them differently: why put some in prison and others in institutions if none of them is responsible?

Empirical evidence will be crucial to answering the various questions posed by the different theories of responsibility. But no amount of scientific evidence on its own will tell us which theory is the right one. This is a theoretical question. Furthermore, as we think about these cases, we should remember that trying to understand the conditions under which people are morally responsible is, in part, a moral problem. We are assessing whether to think of fellow human beings as members of the moral community, as patients, or as objects. In doing so, we should not lose sight of the fact that we are engaged in a theoretical enterprise with a moral dimension.

Taking Stock

We've seen that the question, "if the world is determined, can there be free will?" is not quite the right question to ask. First, the world is likely not completely determined. Second, the kind of indeterminism we seem to have doesn't necessarily help the cause of free will. And third, the real question is whether free will and moral responsibility are compatible with our actions being caused by events that are beyond our control. The theories we have considered answer this question in different ways and they all have their pros and cons. Debate about this topic is very much alive and will likely be kept that way by new scientific discoveries. Whatever decisions we make, we should take care to notice the ways in which our beliefs about free will, moral responsibility, praise, blame, and punishment have serious repercussions for what we do and what kind of society we create.

Summary

- Determinism is the view that the laws of nature and the events of the past entail every truth about the future.
- Compatibilist theories take determinism to be compatible with free will and moral responsibility. Compatibilists hold that free actions for which we are morally responsible are deterministically caused in a certain way (for instance, by one's real self or by one's capacity to grasp and act for reasons). Incompatibilist theories reject the compatibilist thesis.
- There are two main arguments against compatibilism:
    - The Manipulation Argument, which says that if manipulated action is not free, then neither is any action. The Zygote Argument is one prominent version of the Manipulation Argument. Look to the theories discussed in Chapter 10 to see how they distinguish between manipulated action and free action.
    - The Forking Paths Argument, which says that freedom requires genuine alternatives that compatibilism cannot provide. Compatibilists answer this argument by pointing out that control is what's important and distinguishing regulative control (which requires genuine alternatives) from guidance control (which does not require genuine alternatives).
- We can begin to answer the question, "Is free will incompatible with determinism because it requires forking paths?" by thinking about our intuitions about cases.
- By finding out the intuitions of ordinary people, experimental philosophy adds to this endeavor by showing that compatibilism may not be completely at odds with common sense.

- Experimental philosophy doesn't solve the problem by itself, however. Intuitions about cases are relevant to defining a notion of free will, but there are other factors, such as compatibility with scientific evidence and the cost of abandoning moral responsibility.
- There are two kinds of incompatibilism. Hard incompatibilism (or moral responsibility skepticism) accepts determinism and rejects moral responsibility. Libertarianism rejects determinism and thinks we are morally responsible.
- Neuroscientists have argued that free will doesn't exist on the basis of their investigations of brain activity, but these findings do not present a serious challenge to free will according to most philosophers.
- New information from neuroscience and psychology will raise new questions about when people are and are not responsible, questions whose answers will require the resources of moral philosophy.

Study Questions

1. How do you think our society would change if most people became convinced of moral responsibility skepticism and gave up the idea of retributive punishment?
2. I once met a philosopher who believed in hard incompatibilism who said that the most troublesome thing about taking this position was that his girlfriend was upset about it. Why would she be upset? Do you think she should be?
3. Are there some scientific findings that would change your views about free will and responsibility? What kind of experiments would you like to see on these topics?
4. What should theorists do if intuitions about moral responsibility are quite varied, that is, if they change from person to person or within persons from case to case?
5. Research has shown that executive function in the male brain does not really mature until about 25 years of age. Younger men do not have as much brain capacity for impulse control or self-regulation. Should we treat men aged 18 to 25 differently from the way in which we treat men over 25? How?

Notes

[1] Note that Fischer and Ravizza characterize their view as "semi-compatibilism," because they hold that we do have moral responsibility, but not free will. This is a detail we can ignore for our purposes.

[2] Another way comes from Derk Pereboom (2001), whose version of the Manipulation Argument is called "the four case argument" because it uses four progressively difficult cases for the compatibilist.

3 Seventy-six percent of participants thought Fred and Barney acted freely. Sixty percent thought Fred was morally responsible for stealing, and sixty-four percent thought Barney was morally responsible for returning the wallet. In response to scenarios using two more ways of describing determinism, the majority of participants also responded that agents in these scenarios had free will and were morally responsible for their actions (Nahmias et al. 2007).

4 And, of course, there are those who are skeptical of the methods of experimental philosophy altogether (Kauppinen 2007).

5 Agent-causal libertarianism does have defenders, however, who argue that it is the best way to capture the intuitive idea that one's *self* is the cause of one's free actions. See O'Connor 2000.

6 We'll focus on one option here; for more on incompatibilist free will, see Clarke, Capes, & Swenson 2021.

7 For more discussions of Libet's research, see A. L. Roskies (2010a and 2010b) and Sinnott-Armstrong (2014).

8 As mentioned in Chapter 6, it is misleading to talk about *the* psychopath, since in reality people with various capacities fall under the label. In this context, I mean to talk about an extreme case for the purposes of illustrating differences between theories of responsibility.

9 Recall the evidence we saw in Chapter 6 that at least some psychopaths fail to distinguish between moral and conventional wrongs, which has been taken to be evidence that psychopaths do not grasp moral reasons.

Further Readings

Ekstrom, L. 2019. Toward a plausible event-causal indeterminist account of free will. *Synthese 196*(1): 127–144.

Fischer, J. M., R. Kane, D. Pereboom, & M. Vargas. 2007. *Four Views on Free Will*. Blackwell.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences 8*(4): 529–566.

Nahmias, E. 2011. Scientific challenges to free will. In *A Companion to the Philosophy of Action*, T. O'Connor & S. Constantine (eds), pp. 345–356. Wiley.

Nahmias, E., S. G. Morris, T. Nadelhoffer, & J. Turner. 2007. Is incompatibilism intuitive? *Philosophy and Phenomenological Research 73*(1): 28–53.

Nichols, S., & J. Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous 41*(4): 663–685.

Roskies, A. L. 2021. The neuroscience of free will. *U. St. Thomas JL & Pub. Pol'y 15*, 162.

Sinnott-Armstrong, W. (ed.) 2014. *Moral Psychology, Vol. 4: Freedom and Responsibility*. MIT Press.

12 Conclusion

- Is and Ought
- Lessons from Moral Psychology
- Summary
- Study Questions
- Notes

We started out in the Introduction making distinctions between normative (or prescriptive), theoretical (or conceptual), and empirical (or descriptive) questions. As the book has proceeded, we've seen many ways in which these questions and the methods used to answer them are intertwined. It seems appropriate to end with a summary of what we've learned about these connections. The first thing to notice is what we did not see. We did not see a simple derivation of an "ought" from an "is." We evolved to be groupish, but it doesn't follow neatly from the facts of our evolution that we ought to be partial to our ingroup. We have different systems in our brains that give rise to different ideas about what to do if we're confronted with a moral dilemma, but what we actually ought to do doesn't follow simply from the neuroscience without adding some philosophical reflection. The influence of character traits on behavior is not as strong as we may have thought, but it doesn't follow automatically that we should ditch our efforts to cultivate virtue. In every case we've seen, the take-home lesson about the relationship between what is and what ought to be is: it's complicated.

Is and Ought

As complicated as it is, we can make some general observations. Looking back on the chapters of this book, we can distinguish three ways in which the normative, the theoretical, and the empirical are related.

First, descriptive claims about our intuitions can play a role in the justification of normative theories by the method of reflective equilibrium. We've seen several examples of this. For instance, in Chapter 11, we saw that intuitions about whether a person in a deterministic universe is responsible for their actions play a role in an argument for compatibilism. Of course, the argument is

DOI: 10.4324/9781003402787-15

not as simple as "intuitions favor compatibilism, therefore compatibilism is true." Rather, compatibilist intuitions are used against one attack against it, namely, that compatibilism is guilty of changing the subject since it posits such an unintuitive view of responsibility. The fact that people do not in all circumstances find the view unintuitive is just one piece of evidence in favor of compatibilism that must be taken together with all the other relevant information.

We saw another example in Chapter 4 in our discussion of theories of well-being. Nozick's example of the experience machine is an intuition pump designed to make us see the problems with hedonism. This argument against hedonism works (if it does) because the thought experiment brings people to agree that pleasure isn't the only thing worth having in life. If we take this point seriously in reflective equilibrium about the nature of a good human life, we will turn away from hedonism. This example should remind us that what's really relevant to reflective equilibrium are considered judgments or reflective intuitions. Someone who doesn't understand the experience machine example – say, someone who assumes that there couldn't be pleasure without pain and concludes that the experience machine wouldn't be so great after all[1] – doesn't have intuitions that need to be considered in reflective equilibrium. Reflective equilibrium is a method of justification, not a democratic procedure. But the facts about what people would intuit if they understood – facts about people's considered judgments – are still facts that scientific methods could help to uncover. For instance, we could devise tests to see if people understand our thought experiments and focus on the responses of reflective people. If we did this and we discovered that people's considered judgments about cases vary or that people have different intuitions about cases that seem substantially the same, this information is something we should take into account (among other things, of course) in our ethical theorizing.

Because wide reflective equilibrium encompasses background theories, we can identify a second relationship between the descriptive and the normative. Normative theories make empirical assumptions, and the empirical facts are relevant to whether these assumptions are true or false. (The first point about intuitions is really just a special case of this second relationship.) We have seen many examples of this. From Chapter 3, recall that many moral theories hold that true moral motivation is other regarding. For Kantians, actions must be motivated by the sense of duty to count as morally worthy, and for virtue ethicists, ethical action is motivated by the emotional states and concerns that are constitutive of the virtues (such as compassion or friendliness). Now, if people are necessarily egoistic and it is psychologically impossible for us to have any motive that is not directed toward some good for ourselves, then these theories are (at least in this one respect) either incorrect or inapplicable to human beings. Of course, the science of egoism and altruism has probably not shown that it is impossible for us to be motivated by altruistic desires or by duty, but the point here is that this science is relevant to normative theories.

A similar point was made in Chapter 9 on virtue. Virtue ethical theories assume that people can develop virtues. Insofar as virtues are states of our psychology, as Aristotle took them to be, scientific evidence about our psychology is going to be relevant to the assumption that we can develop virtues. As we saw, critics have charged that Aristotelian virtue ethics relies on a picture of virtues that is psychologically implausible because it depicts us as possessing highly stable dispositions that ensure we will act well in any situation. We also saw some ways in which virtue ethicists have clarified or modified their assumptions in response to this charge. Virtue ethics wasn't killed by the situationist critique, but the facts about what kind of traits we actually have are certainly relevant and will need to be taken into consideration as virtue ethics develops. If a theory doesn't apply to human beings unless they have a certain kind of psychological make-up, it's important to know whether human beings are actually made this way and that is a matter of empirical fact.

We also have the example of well-being and the reasons to be moral from Chapter 4. Various moral theories throughout history have assumed that there are self-interested reasons to be moral, and this is an assumption that is open to scientific investigation, once the terms are defined. Once we have answered the philosophical questions of what self-interest is, and what counts as acting morally, it remains to consider whether the one causes the other, and this is an empirical question. Testing these assumptions about the link between morality and self-interest will raise a number of other philosophical questions: What kind of reasons do we have to be moral that stem from well-being? Are these reasons overriding reasons? If not, are these the wrong kind of reasons? These are not questions that can be answered by science, but psychological research is certainly important for assessing the normative conclusions of theories that posit a relationship between self-interest and morality.

Finally, the third relationship we can identify is between the descriptive and the metaethical. Metaethical theories make empirical assumptions too, and scientific evidence is relevant to them insofar as it tests these assumptions or adds new relevant information. The main example we have considered of the way in which empirical research is relevant to metaethics is about the role of sentiment or desire in moral judgment and motivation. The sentimentalist theory of moral judgment assumes that there is an intimate connection between moral judgment and sentiment. Evidence for some connection comes from experiments that show that emotions influence our moral judgments in ways that would be surprising if there were no connection between judgment and sentiment. For example, feeling disgusted by our immediate surroundings (because of fart spray, say) makes us judge wrong-doers more harshly. Further, sentimentalism predicts that people with defective sentiments will also have defects in their capacity for moral judgment. We saw some evidence that this is indeed the case. Psychopaths have an emotional deficit, and they also have trouble making the right distinction between moral and conventional norms: they do not seem to understand the gravity of morality.

Notice, though, that the scientific evidence does not carry the day by itself. If we operationalize moral judgment in terms of the content (the specific facts that these judgments pick out, such as facts about harm) rather than in terms of other features of moral judgment (such as seriousness or authority independence), then psychopaths' capacities for judgment do not seem to be defective. If we start this way, we will probably conclude that while psychopaths know what is moral and what is immoral, they just don't care. In this way our philosophical assumptions about moral judgment influence how we interpret the empirical evidence. Which is a better starting point depends on the big picture and what we take the point of moral judgment to be. I think it makes more sense to think of the point of moral judgment in terms of its role in helping us live together, than to think of it in terms of its role in picking out a particular set of facts. This is not so for judgments about the weather – here we should understand our judgments as responding to facts about rain, temperature, and so on. But when it comes to morality, it seems to me correct that the defining feature is its practical role. This inclines me to take seriously the research on psychopaths who do not know how to live in community with others. The fact that they don't understand the practical importance of moral judgment is a profound defect. But these thoughts are preceded by some philosophical reflection on how it makes sense to think about the whole moral enterprise.

Psychological research about psychopaths and the role of sentiments puts pressure on the Kantian to qualify her claims. It does not refute the entire Kantian project, however. We have seen that the Kantian can clarify or modify the picture to preserve the basic insights in light of the research on the role of emotions in moral judgments. In particular, the Kantian can preserve the basic idea that correct moral judgments are justified by rational principles, but reject the idea that moral judgments are typically *caused* by reasoning. As long as we are, at our best and most rational, capable of investigating whether our moral judgments have a principled basis and capable of rejecting the ones that do not, the Kantian could be pretty happy. Greene's evidence about the emotional basis for deontological judgments in trolley cases (discussed in Chapter 8) adds some fuel to the fire, because it makes it seem that emotions are playing a differentially strong role in principled Kantian judgments as opposed to consequentialist judgments. But here again if the Kantian distances herself from claims about the causes of moral judgment and insists that her focus is on justification, she can argue that which moral judgments are justified by rational principles (the real question) is independent of what parts of the brain are activated when "characteristically deontological" judgments are made. The Kantian may even want to reject some "characteristically deontological" judgments on the basis of Greene's research. A Kantian who was sympathetic to Greene's project might say that some of his findings shed light on what courses of action are really justified by the Categorical Imperative, by demonstrating that some previous conclusions were the result of irrational, emotional bias rather than objective application of a principle.

Kantians can defend themselves against attacks based on research about the causes of moral judgments, but there is another line of attack, which we considered briefly in Chapter 7 in our discussion of the Kantian challenge to sophisticated sentimentalism. The challenge, in particular as it is formulated by

Korsgaard, is that given the kind of reflective creature a human being is, we cannot make do with sentiments when we are looking for reasons: sentiments (or desires, for that matter) do not justify. In other words, reflective creatures like us need to find something at the bottom of the pile of reasons that puts an end to our questions, something like a purely rational principle such as the Categorical Imperative. This argument assumes something about our psychology, namely, that we are reflective creatures of a particular kind: ones whose conviction that we have reasons to do anything at all rests fundamentally on the idea that there are universal principles to support these reasons. But what if instead of being reflective creatures of this kind, our moral convictions are entirely unshaken by the thought that there are no universal principles to underwrite them? What if we are perfectly happy to acknowledge that if it weren't for our having the desires or the sentiments we have, nothing would be either moral or immoral? This would be a problem for the Kantian argument under consideration. The fact that we don't need rational principles (if it is a fact) would deflate the argument that there must be such principles if there are to be any reasons at all. And this would be a case in which the facts about our psychology are relevant to the assumptions made by a rationalist metaethical theory.[2]

We see that when it comes to metaethics, the relationship between philosophical questions and scientific research is complex. Metaethical questions are questions about our moral practice and the normative theories that are part of it – what moral judgments and theories are about, how these judgments and theories may be justified, and so on. The concepts involved in these questions (e.g., reason, fact, justification) are difficult on their own and intertwined in ways that make it even more difficult to know exactly how they should be understood. Answering metaethical questions requires paying attention to science and paying attention to conceptual subtleties until we reach a point at which our concepts help us make the best sense of all the information we have.

So, even if you can't derive an ought from an is, the facts about our psychology do matter for ethics.

Lessons from Moral Psychology

I want to conclude with three of the most important things I think we learn from research in moral psychology. These are not the only important lessons, to be sure, but they are three that I think are worth highlighting because they matter to our everyday lives.

First, we are emotional creatures and our emotions are important to ethics in a way philosophers have sometimes failed to notice. I believe psychology is showing us that that David Hume was right at least about one thing: without emotions, we would not have morality at all. We would not care about each other, we would not value anything, and we would not worry about making good, defensible decisions about what to do. If this is true about us, there are some implications for metaethics – for what moral judgments are and whether they are inherently motivating (although, as we've discussed, these implications have sometimes been exaggerated). There are also implications for what is good for us. There has been a tendency in the history of philosophy and religion to think of the emotional self as a beast in need of taming. Life goes best for us when our reason rules the beast and brings it into line. But if emotions are what allow us to have values in the first place, then emotions are not necessarily unruly beasts; they are also sources of information about ourselves and the world. Acknowledging this – elevating the emotions to "partner" status – should make us think differently about what a good human life is. Perhaps the aim should not be to control our emotions, but to live in harmony with them. Living in harmony with our emotional selves does not mean giving up on reflection and thinking, but it does mean that when we are reflecting and thinking, we should listen to what our hearts have to tell us. Not everything that comes from the heart is worth heeding, but then not everything that comes from the head is worth heeding either.

Second, we have less rational control than we have often assumed, but we do have some. Psychology is showing us that we are often caused to do things by situational factors that bypass our rational capacities, but this does not mean that there are no virtues, because there are ways of understanding virtue that are compatible with these facts. Nor do I think the research shows that we are never morally responsible, because I believe that we have enough of a certain kind of control to make sense of attributions of responsibility. We can, for example, make long-term plans and adopt strategies for coping with momentary temptations that will help us to act consistently with our ultimate desires and our better judgment. We can also use our rational capacities to reach better moral conclusions. We can reason by analogy to extend our moral sympathies more broadly, for example, and we can use awareness of our biases to reconsider some of our knee-jerk moral reactions.

I do think, however, that the fact that our rational capacities are not the captains of our ships should make us kinder to each other and to ourselves. We are often harsh judges: we get angry at people for slighting us, down on ourselves for slipping up on diets or exercise plans, unforgiving of those who have done something wrong, deeply ashamed of our own faults. The picture behind all this negative attitude seems to be that there is a pure agent inside each of us who could be perfect if only he or she would exert a little willpower! But this isn't how we are. Insofar as psychology is helping us to see that this isn't how we are, it provides an important lesson, namely, that we should ease up a little bit.

At the same time, there is a certain kind of judgment and self-criticism that is highly appropriate, given what we know about our psychology. If we are strongly influenced by forces that we do not sanction as *good* influences, then some of what we think, want, and do is probably not terribly well justified. Some of the things we were raised to believe about what's morally right and wrong have nothing else going for them than the fact that we were raised that way. Some of the moral judgments we make in the moment may be more the result of what we had for breakfast than the result of any moral insight.[3] Understanding this about our psychology should lead us to have a little humility about our own moral perspectives.

Finally, we are profoundly social creatures. We saw this very clearly in Part I of the book, in discussions of the starting points of morality and the relationship between self-interest and morality. Throughout evolution, and from birth, human beings are inclined to care about others. We have also seen it in our discussions of emotions that are highly attuned to the actions of other people, and in our discussions about responsibility and blame. Our practices of holding each other responsible are essentially social practices that we need in order to work together and thrive. We would expect isolation to be bad for such social creatures – and, indeed, the Covid pandemic provided a massive natural experiment that confirmed this expectation. As we go about thinking about how to live our lives and how to structure our societies, we would do well to remember how much of our individual psychology is fitted to our need to live and work together.

Those are some lessons that I value; there are many more. In the conclusion of the first edition of this book I expressed the hope that, as the field develops, collaboration between psychologists and philosophers would increase, producing new discoveries and ideas and an even better understanding of why we are good when we are good, bad when we are bad, and what it means to make these judgments. This has certainly turned out to be the case. Indeed, interdisciplinary moral psychology has expanded far beyond my hopes.

Summary

- Normative, theoretical, and empirical questions and methods are intertwined in many complex and interesting ways.
- First, facts about people's intuitions are relevant to reflective equilibrium justifications of normative theories (e.g., free will and intuitions about compatibilism; intuitions about happiness and the experience machine).
- Second, scientific facts are relevant to the empirical assumptions made by normative theories (e.g., altruism, virtue, Kantian reflectivism, self-interested reasons to be moral).
- Third, scientific facts are relevant to the empirical assumptions made by metaethical theories (e.g., moral judgment sentimentalism vs. rationalism), which may ultimately have implications for normative theory.

Study Questions

1. Can you think of an example (from your own life, or from a film or novel) in which someone would have been better off listening to their emotions rather than following their reasoning? And an example where it went the other way?

2. Is there anything you have learned about human psychology that has changed how you think about your life (for example, the meaning of your life, how you ought to live your life, or how you ought to treat others)?
3. What do you think is the most important thing psychology has to teach moral philosophers?
4. What do psychologists have to learn from moral philosophy?

Notes

[1] This would represent a failure to understand the example, because the experience machine by hypothesis guarantees more pleasure overall. Therefore, if it's true that you need to have some pain in order to have pleasure, the machine would make sure that you have just enough pain so that you can get the most pleasure possible.

[2] Things in this case are quite tricky, and I can imagine the Kantian denying that the empirical facts are relevant here. What is relevant, she might say, is what we are like insofar as we are rational and that is not a psychological fact about us. This is a fair point, and the Kantian argument is certainly in part conceptual: the relevant premise is that a consideration cannot count as a reason unless it would be sanctioned by a rational law – otherwise it is always sensible to ask whether it really is a reason to do something or not. That said, if we cannot recognize ourselves (even our best, most rational selves) in the description of rational agents in this argument, then it's unclear what reason we have to care about reasons as Kantians see them. This by itself would not prove that the Kantian conception of a reason is wrong, exactly, but it would make us question what the point of it is.

[3] Even judges in a court of law are susceptible to the influence of extraneous factors, according to Danziger, Levav, and Avnaim-Pesso (2011), who found that judges are more lenient after a break for lunch.

Glossary of Theories and Technical Terms

Categorical Reasons Reasons that apply to us independently of our desires. Moral reasons are often taken to be categorical. Universal reasons, by contrast, are reasons that apply to everyone; they may or may not be contingent on our desires.

Causal Determinism The thesis that every event is necessitated by antecedent events and conditions together with the laws of nature. Or, to put it another way, the facts of the past, in conjunction with the laws of nature, entail every fact about the future.

Compatibilism and Incompatibilism *Compatibilism:* if determinism is true, we could have free will. *Incompatibilism*: if determinism is true, we do not have free will. *Hard incompatibilism*: whether or not determinism is true, we do not have free will. *Libertarianism*: determinism is false and we do have free will (a form of incompatibilism).

Construct Validity In psychology, a measurement tool (such as a survey or a test) has construct validity when it measures what it is supposed to measure (the real, ultimate aim of investigation).

Cybernetics The study of principles governing goal-directed systems that self-regulate via feedback. Also called "control theory." From the cybernetic perspective, a desire is a representation of the state toward or away from which the cybernetic system moves.

Dual Process Theory The theory that there are two different cognitive systems that we use for many purposes, such as making a moral judgment: System 1 is fast, automatic, and non-conscious; System 2 is slow, controlled and conscious.

Eudaimonism According to Aristotelian eudaimonism, a good life for a person is one in which she fulfills her *human nature* (her nature as a member of the human species). Individualist eudaimonism says that a good life is one in which a person fulfills her individual nature.

Hedonism Hedonism about well-being is the view that well-being consists in pleasure and the absence of pain.

Humean Theory of Motivation The theory that desires are necessary for motivating actions; beliefs never motivate by themselves.

Intentional Content Roughly, "aboutness": a mental state has intentional content if it is directed onto something in the world. Beliefs, desires, and emotions have intentional content. Intentional content is not the same thing as a person's intentions.

Internalisms and Externalisms *Moral Judgment Internalism*: the view that, in making a moral judgment, one is thereby motivated, to some degree, to act on it. *Moral Judgment Externalism*: the view that in making a moral judgment one is not thereby motivated. *Reasons Existence Internalism* (RI) the view that normative reasons are necessarily motivating, at least under certain conditions. *Reasons Existence Externalism*: the view that normative reasons are not necessarily motivating.

Moral and Conventional Norms Moral norms are thought to be more serious and have wider applicability than conventional norms. Conventional norms are thought to be contingent on an authority (such as a teacher or the law), and they receive a different kind of justification from moral norms, which are often justified in terms of harm or fairness.

Moral Intuition A moral judgment that appears to be fairly obvious to you without argument or inference.

Moral Nativism The view that there are moral dispositions prior to learning.

Motivating Reasons and Normative Reasons Motivating reasons explain actions; normative reasons are considerations that count in favor of and justify actions.

Practical Reasons and Theoretical Reasons Practical reasons are reasons for action; theoretical reasons are reasons for belief.

Psychological Egoism The theory that all voluntary action is selfish or produced by self-interested desires.

Rationalism Moral Rationalism (in the context of this book) is the view that the truth of a moral judgment is determined by rational principles. Moral judgments are justified and give us normative reasons for action insofar as they conform to these principles.

Reflective Equilibrium The predominant method for constructing and defending ethical theories, which proceeds by bringing into a coherent whole considered judgments (or intuitions) about cases, putative principles, and background theories.

Retributivism The view that punishment is justified by the fact that a person who performs certain kinds of actions (such as actions that harm other innocent people) *deserves* to be punished. It contrasts with forward-looking or consequentialist justifications of punishment.

Sentimentalism The view that moral judgments are expressions of, or reports about, our sentiments. Sentimentalists reject the idea (held by rationalists) that moral judgments are justified by rational principles; they require a different explanation of the fact that we do justify our moral judgments by appeal to reasons. Sophisticated sentimentalists aim to provide this explanation.

Bibliography

Ahadi, S., & E. Diener. 1989. Multiple determinants and effect size. *Journal of Personality and Social Psychology 56*(3): 398–406.

Aharoni, E., W. Sinnott-Armstrong, & K. A. Kiehl. 2012. Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology 121*(2): 484–497.

Aknin, L. B., C. P. Barrington-Leigh, E. W. Dunn, J. F. Helliwell, J. Burns, R. Biswas-Diener … & M. I. Norton. 2013. Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology 104*(4): 635.

Aknin, L. B., E. W. Dunn, & A. V. Whillans. 2022. The emotional rewards of prosocial spending are robust and replicable in large samples. *Current Directions in Psychological Science 0*(0). https://doi.org/10.1177/09637214221121100

Aknin, L. B., A. V. Whillans, M. I. Norton, & E. W. Dunn. 2019. Happiness and prosocial behavior: An evaluation of the evidence. *World Happiness Report 2019*, 67–86.

Alfano, M. 2013. *Character as Moral Fiction*. Cambridge University Press.

Altham, J. E. J. 1986. The legacy of emotivism. In Fact, Science and Morality: Essays on A. J. Ayer's Language, Truth and Logic, G. Macdonald & C. Wright (eds), pp. 275–288. Blackwell.

Anderson, E. 1995. *Value in Ethics and Economics*. Harvard University Press.

Andrews, K., & L. Gruen. 2014. Empathy in other apes. In *Empathy and Morality*, H. Maibom (ed.), pp. 193–209. Oxford University Press.

Anscombe, G. E. M. 1957. *Intention*. Harvard University Press.

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy 33*(124): 1–19.

Antony, L. M. 2000. Natures and norms. *Ethics 111*(1): 8–36.

Appiah, K. A. 2008. *Experiments in Ethics*. Harvard University Press.

Aristotle. 1985. *Nicomachean Ethics*, T. Irwin (trans.). Hackett.

Arpaly, N. 2000. Hamlet and the utilitarians. *Philosophical Studies 99*(1): 45–57.

Arpaly, N. 2002. Moral worth. *The Journal of Philosophy 99*(5): 223–245.

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford University Press.

Arpaly, N., & T. Schroeder. 1999. Praise, blame and the whole self. *Philosophical Studies 93*(2): 161–188.

Arpaly, N., & T. Schroeder. 2014. *In Praise of Desire*. Oxford University Press.

Aslin, R. N. 2007. What's in a look?. *Developmental Science 10*(1): 48–53.

Austen, J. 2000. *Emma*. 3rd edition. W. W. Norton & Company.

Ayer, A. J. 1952. *Language, Truth, and Logic*. Dover.

Aznar, A., H. R. Tenenbaum, & P. S. Russell. 2022. Is moral disgust socially learned? *Emotion*, Feb. 7. doi: 10.1037/emo0001066

Bago, B., B. Aczel, Z. Kekecs, J. Protzko, M. Kovacs, T. Nagy, T. Gill, U.-D. Reips … & C. R. Chartier. 2020. (Preregistered and conditionally accepted). Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgments. *Nature Human Behaviour*.

Bago, B., & W. De Neys. 2017. Fast logic?: Examining the time course assumption of dual process theory. *Cognition 158*, 90–109.

Baron, M. 1999. *Kantian Ethics Almost without Apology*. Cornell University Press.

Barrett, L. F. 2017. *How Emotions Are Made. The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Batson, C. D. 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Lawrence Erlbaum.

Batson, C. D., B. D. Duncan, P. Ackerman, T. Buckley, & K. Birch. 1981. Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology 40*(2): 290–302.

Baumeister, R. F. 2022. Self-regulation and conscientiousness. In *Noba Textbook Series: Psychology*, R. Biswas-Diener & E. Diener (eds), DEF publishers. URL = http://noba.to/3j96qxwr

Baumeister, R. F., E. J. Masicampo, & C. N. DeWall. 2009. Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and social psychology bulletin 35*(2): 260–268.

Baumeister, R. F., & M. R. Leary. 1995. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin 117*(3): 497–529.

Berker, S. 2009. The normative insignificance of neuroscience. *Philosophy & Public Affairs 37*(4): 293–329.

Berridge, K. C. 2003. Pleasures of the brain. *Brain and Cognition 52*(1): 106–128.

Besser-Jones, L. 2008. Personal integrity, morality and psychological well-being: Justifying the demands of morality. *The Journal of Moral Philosophy 5*: 361–383.

Besser-Jones, L. 2012. The role of practical reason in an empirically informed moral theory. *Ethical Theory and Moral Practice 15*(2): 203–220.

Besser, L. 2017. Virtue of self-regulation. *Ethical Theory and Moral Practice 20*(3): 505–517.

Besser, L. 2022. Virtue. In *Oxford Handbook of Moral Psychology*, M. Vargas & J. Doris (eds). Oxford University Press.

Bjorklund, F., J. Haidt, & S. Murphy. 2000. *Moral dumbfounding: When intuition finds no reason*. Unpublished manuscript, University of Virginia.

Blackburn, S. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. Clarendon Press.

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Clarendon Press.

Blair, R. J. R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition 57*(1): 1–29.

Bloom, P. 2017. *Against Empathy: The Case for Rational Compassion*. Random House.

Bloomfield, P. 2014. *The Virtues of Happiness: A Theory of the Good Life*. Oxford University Press.

Bowles, S., & H. Gintis. 2011. A cooperative species. In *A Cooperative Species*. Princeton University Press.

Bramble, B. 2013. The distinctive feeling theory of pleasure. *Philosophical Studies 162*(2): 201–217.

Bramble, B. 2016. A new defense of hedonism about well-being. *Ergo, an Open Access Journal of Philosophy 3*.

Brink, D. O., & D. K. Nelkin. 2013. Fairness and the architecture of responsibility. *Oxford Studies in Agency and Responsibility 1*, 284–313.

Cameron, C. D., P. Conway, & J. A. Scheffer. 2022. Empathy regulation, prosociality, and moral judgment. *Current Opinion in Psychology 44*, 188–195.

Carey, J. M., & D. L. Paulhus. 2013. Worldview implications of believing in free will and/or determinism: Politics, morality, and punitiveness. *Journal of Personality 81*(2): 130–141.

Carver, C., & M. Scheier. 1998. *On the Self-Regulation of Behavior*. Cambridge University Press.

Caspar, E. A., L. Vuillaume, P. A. Magalhães De Saldanha da Gama, & A. Cleeremans. 2017. The influence of (dis) belief in free will on immoral behavior. *Frontiers in Psychology 8*: 20.

Chappell, T. 2014. *Knowing What To Do: Imagination, Virtue, and Platonism in Ethics*. Oxford University Press.

Cherry, M. 2021. *The Case for Rage: Why Anger Is Essential to Anti-Racist Struggle*. Oxford University Press.

Churchland, P. S. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton University Press.

Ciurria, M. 2014. Answering the situationist challenge: A defense of virtue ethics as preferable to other ethical theories. *Dialogue: Canadian Philosophical Review 53*(4): 651–670.

Ciurria, M. 2019. An *Intersectional Feminist Theory of Moral Responsibility*. Routledge.

Clarke, R., J. Capes, & P. Swenson. 2021. Incompatibilist (nondeterministic) theories of free will. *The Stanford Encyclopedia of Philosophy* (Fall 2021 edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories/

Cohen, J. 1988. The effect size. *Statistical Power Analysis for the Behavioral Sciences*, pp. 77–83. Erlbaum.

Crisp, R. 2006. Hedonism reconsidered. *Philosophy and Phenomenological Research 73*(3): 619–645.

Cushman, F. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review 17*(3): 273–292.

Dahl, A. 2019. The science of early moral development: On defining, constructing, and studying morality from birth. *Advances in Child Development and Behavior 56*: 1–35.

Dahl, A., C. P. Baxley, & T. Waltzer. 2021. The two-front forever war: Moral nativism and its critics. *Human Development 65*(3): 180–187.

Dancy, J. 2000. *Practical Reality*. Oxford University Press.

Daniels, N. 1979. Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy 76*(5): 256–282.

Daniels, N. 2008. Reflective equilibrium. *Stanford Encyclopedia of Philosophy*. www.illc.uva.nl/~seop/archives/fall2008/entries/reflective-equilibrium/

Danziger, S., J. Levav, & L. Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences 108*(17): 6889–6892.

Darley, J. M., & C. D. Batson. 1973. "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology 27*(1): 100–108.

D'Arms, J., & D. Jacobson. 2000. Sentiment and value. *Ethics 110*(4): 722–748.

Deci, E. L., & R. M. Ryan. 2004. *Handbook of Self-Determination Research*. University of Rochester Press.

De Dreu, C. K. W. 2012. Oxytocin modulates cooperation within and competition between groups: An integrative review and research agenda. *Hormones and Behavior 61*(3): 419–428.

De Dreu, C. K. W., L. L. Greer, M. J. J. Handgraaf, S. Shalvi, G. A. Van Kleef, M. Baas, F. S. Ten Velden, E. Van Dijk, & S. W. W. Feith. 2010. The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science 328*(5984): 1408–1411.

Deery, O., & E. Nahmias. 2017. Defeating manipulation arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies 174*(5): 1255–1276.

Dennett, D. C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.

Dennett, D. C. 2004. *Freedom Evolves*. Penguin.

Deonna, Julien A., & F. Teroni. 2012. *The Emotions: A Philosophical Introduction*. Routledge.

Deonna, Julien A., & F. Teroni. 2014. In what sense are emotions evaluations? In *Emotion and Value*, Sabine Roeser & Cain Todd (eds), pp. 15–31. Oxford University Press.

Deonna, Julien A., & F. Teroni. 2016. Getting bodily feelings into emotional experience in the right way. *Emotion Review 9*(1): 55–63.

De Waal, F. B. 1997. *Good Natured: The Origins of Right and Wrong In Humans and Other Animals*. Harvard University Press.

DeYoung, C. G. 2015. Cybernetic big five theory. *Journal of Research in Personality 56*, 33–58.

DeYoung, C. G., & V. Tiberius. 2022. Value fulfillment from a cybernetic perspective: A new psychological theory of well-being. *Personality and Social Psychology Review 7*(1): 3–27. doi: 10.1177/10888683221083777. Epub 2022 Apr. 20.

DeYoung, C. G., & Y. J. Weisberg. 2019. Cybernetic approaches to personality and social behavior. In *Oxford Handbook of Personality and Social Psychology* (2nd edn), M. Snyder & K. Deaux (eds), pp. 387–414. Oxford University Press.

Dobbin, F., & A. Kalev. 2016. Why diversity programs fail and what works better. *Harvard Business Review 94*(7–8): 52–60.

Dominguez, N. 2020. Moral responsibility for implicit biases: Examining our options. In *An Introduction to Implicit Bias*, pp. 153–173. Routledge.

Doris, J. M. 1998. Persons, situations, and virtue ethics. *Nous 32*(4): 504–530.

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge University Press.

Doris, J. M. 2009. Skepticism about persons. *Philosophical Issues 19*: 57–91.

Doris, J. M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Doris, J. M. 2022. *Character Trouble: Undisciplined Essays on Moral Agency and Personality*. Oxford University Press.

Doris, J. M., & the Moral Psychology Research Group. 2010. *The Moral Psychology Handbook*. Oxford University Press.

Downes, Stephen M. 2021. Evolutionary psychology. *The Stanford Encyclopedia of Philosophy* (Spring edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/evolutionary-psychology/

Driver, J. 2001. *Uneasy Virtue*. Cambridge University Press.

Dunn, E. W., L. B. Aknin, & M. I. Norton. 2008. Spending money on others promotes happiness. *Science 319*(5870): 1687–1688.

Dupre, J. 2012. Against maladaptationism: Or, what's wrong with evolutionary psychology? In J. Dupre, *Processes of Life: Essays in Philosophy of Biology*, pp. 245–260. Oxford University Press.

Edmonds, G. W., J. J. Jackson, J. V. Fayard, & B. W. Roberts. 2008. Is character fate, or is there hope to change my personality yet? *Social and Personality Psychology Compass 2*(1): 399–413.

Eisenberg, N. 2000. Emotion, regulation, and moral development. *Annual Review of Psychology 51*(1): 665–697.

Ekman, P., & W. V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology 17*(2): 124–129.

Ekstrom, L. 2019. Toward a plausible event-causal indeterminist account of free will. *Synthèse 196*: 127–144.

Elliott, C. 2022. The degradation drug. *The American Scholar*, September 29, 2022. URL = https://theamericanscholar.org/the-degradation-drug/

Feinberg, J. 2004. Psychological egoism. In *Reason and Responsibility*, R. Shafter-Landau & J. Feinberg (eds), pp. 476–488. Wadsworth.

Feldman, F. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties and Plausibility of Hedonism*. Oxford University Press.

Finlay, S., & M. Schroeder. 2008. Reasons for action: Internal vs. external. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2008/entries/reasons-internal-external/

Fischer, John Martin. 1982. Responsibility and control. *Journal of Philosophy 79*(1): 24–40.

Fischer, J. M. 2007. Compatibilism. In *Four Views on Free Will*, J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (eds), pp. 44–84. Blackwell.

Fischer, J. M., R. Kane, D. Pereboom, & M. Vargas. 2007. *Four Views on Free Will*. Blackwell.

Fischer, J. M., & M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Fischer, P., J. I. Krueger, T. Greitemeyer, C. Vogrincic, A. Kastenmüller, D. Frey … & M. Kainbacher. 2011. The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin 137*(4): 517.

FitzGerald, C., & S. Hurst. 2017. Implicit bias in healthcare professionals: A systematic review. *BMC Medical Ethics 18*(1): 1–18.

Fleeson, W. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology 80*(6): 1011–1027.

Foot, P. 1972. Morality as a system of hypothetical imperatives. *The Philosophical Review 81*(3): 305–316.

Foot, P. 2001. *Natural Goodness*. Oxford University Press.

Foot, P. 2002/1967. The problem of abortion and the doctrine of double effect. *Reprinted in Virtues and Vices and Other Essays in Moral Philosophy*, pp. 19–32. Oxford University Press.

Frankfurt, Harry G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy 66*(23): 829–839.

Frankfurt, H. G. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy 68*(1): 5–20.

Frankfurt, H. G. 1988. *The Importance of What We Care about: Philosophical Essays*. Cambridge University Press.

Frankish, K. 2010. Dual-process and dual-system theories of reasoning. *Philosophy Compass 5*(10): 914–926.

Frijda, Nico H. 1986. *The Emotions*. Cambridge University Press.

Gauthier, D. P. 1986. *Morals by Agreement*. Oxford University Press.

Gibbard, A. 1992. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Oxford University Press.

Gibbard, A. 2006. Moral Feelings and Moral Concepts. *Oxford Studies in Metaethics 1*: 195–215.

Goldie, P. 2000. *The Emotions: A Philosophical Exploration*. Oxford University Press.

Goldie, P. 2007. Emotion. *Philosophy Compass 2*(6): 928–938.

Goldman, A. 2012. A liberal learns to compete. *The New York Times Magazine*, July 27. www.nytimes.com/2012/07/29/magazine/a-liberal-learns-to-compete.html

Grace-Martin, K. Two types of effect size statistic: Standardized and unstandardized. *The Analysis Factor*. https://www.theanalysisfactor.com/two-types-effect-size-statistic/

Graham, J., J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, & P. H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology* (Vol. *47*), pp. 55–130. Academic Press.

Greene, J. 2014a. Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics 124*(4): 695–726.

Greene, J. 2014b. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin Press.

Greene, J. 2023. Trolleyology: What it is, why it matters, what it's taught us, and how it's been misunderstood. In H. Lillehammer, *The Trolley Problem*, pp. 158–181. Cambridge University Press.

Greene, J., F. Cushman, L. Stewart, K. Lowenberg, L. Nystrom, & J. Cohen. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition 111*(3): 364–371.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, & J. D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537): 2105–2108.

Griffiths, P. E. 2002. What is innateness? *The Monist 85*(1): 70–85.

Griffiths, Paul E., Edouard Machery, & Stefan Linquist. 2009. The vernacular concept of innateness. *Mind and Language 24*(5): 605–630.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review 108*(4): 814–834.

Haidt, J. 2007. The new synthesis in moral psychology? *Science 316*(5827): 998–1002.

Haidt, J. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Penguin.

Haidt, J., & J. Baron. 1996. Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology 26*(2): 201–218.

Haidt, J., & F. Bjorklund. 2008. Social intuitionists answer six questions about morality. In *Moral Psychology*, vol. *2*, W. Sinnott-Armstrong (ed.), pp. 181–217. MIT Press.

Halberstadt, J. B., & T. Wilson. 2008. Reflections on conscious reflection: Mechanisms of impairment by reasons analysis. In *Reasoning: Studies of Human Inference and Its Foundations*, J. Adler and L. Rips (eds), 548–565. Cambridge University Press.

Hamlin, J. K., K. Wynn, & P. Bloom. 2007. Social evaluation by preverbal infants. *Nature 450*(7169): 557–559.

Hannikainen, I. R., E. Machery, D. Rose, S. Stich, C. Y. Olivola, P. Sousa, … & J. Zhu. 2019. For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology*, 2428. doi:10.3389/fpsyg.2019.02428. Corpus ID: 207759263

Hare, R. D., & H. Vertommen. 2003. *The Hare Psychopathy Checklist – Revised*. Multi-Health Systems.

Harman, G. 1999. Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society 99*: 315–331.

Harris, S. 2011. *The Moral Landscape: How Science Can Determine Human Values*. Simon and Schuster.

Haybron, D. M. 2008. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. Oxford University Press.

Heathwood, C. 2007. The reduction of sensory pleasure to desire. *Philosophical Studies 133*(1): 23–44.

Heathwood, C. 2011. Preferentism and self-sacrifice. *Pacific Philosophical Quarterly 92*(1): 18–38.

Helm, Bennett W. 2001. *Emotional Reason: Deliberation, Motivation and the Nature of Value*. Cambridge University Press.

Henrich, J. 2016. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press.

Henrich, J., S. J. Heine, & A. Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences 33*(2–3): 61–83.

Herman, B. 1981. On the value of acting from the motive of duty. *The Philosophical Review 90*(3): 359–382.

Hieronymi, P. 2020. *Freedom, Resentment, and the Metaphysics of Morals*. Princeton University Press.

Hill, T. E. 1992. *Dignity and Practical Reason in Kant's Moral Theory*. Cornell University Press.

Hill, T. E. 2019. Making exceptions without abandoning the principle: Or how a Kantian might think about terrorism. In *Dignity and Practical Reason in Kant's Moral Theory*, pp. 196–225. Cornell University Press.

Hirstein, W., K. L. Sifferd, & T. K. Fagan. 2018. *Responsible Brains: Neuroscience, Law, and Human Culpability*. MIT Press.

Hobbes, T. 1994/1651. *Leviathan: With Selected Variants from the Latin Edition of 1668*. Edwin Curley (ed.). Hackett.

Holroyd, J., R. Scaife, & T. Stafford. 2017. Responsibility for implicit bias. *Philosophy Compass 12*(3): e12410.

Hooker, B. 2002. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford University Press.

Horne, Z., D. Powell, & J. Hummel. 2015. A single counterexample leads to moral belief revision. *Cognitive Science 39*: 1950–1960.

Hrdy, S. B. 2009. *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Harvard University Press.

Huang, K., J. D. Greene, & M. Bazerman. 2019. Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences 116*(48): 23989–23995.

Hume, D. 2000/1739. *A Treatise of Human Nature*. D. F. Norton and M. J. Norton (eds). Oxford University Press.

Hursthouse, R. 1999. *On Virtue Ethics*. Oxford University Press.

Hursthouse, Rosalind, & Glen Pettigrove. (Winter 2022 Edition). Virtue ethics. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds), https://plato.stanford.edu/archives/win2022/entries/ethics-virtue/

Isen, A. M., & P. F. Levin. 1972. Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology 21*(3): 384–388.

Jacobson, D. 2011. Fitting attitude theories of value. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). http://plato.stanford.edu/archives/spr2011/entries/fitting-attitude-theories/

James, W. 1884. What is an emotion? *Mind 9*: 188–205.

Jayawickreme, E., Peter Meindl, E. G. Helzer, R. M. Furr, & W. Fleeson. 2014. (Invited and under review). Virtuous states and virtuous traits: How the empirical evidence regarding the existence of broad traits does not undermine virtue ethics. *Theory and Research in Education 12*(3).

John, O. P., R. W. Robins, & L. A. Pervin. 2008. *Handbook of Personality: Theory and Research*. Guilford Press.

Johnson, R. N. 1999. Internal reasons and the conditional fallacy. *The Philosophical Quarterly 49*(194): 53–72.

Jones, E. E., & V. A. Harris. 1967. The attribution of attitudes. *Journal of Experimental Social Psychology 3*(1): 1–24.

Jones, K. 2006. Metaethics and emotions research: A response to Prinz. *Philosophical Explorations 9*(1): 45–53.

Kamtekar, R. 2004. Situationism and virtue ethics on the content of our character. *Ethics 114*(3): 458–491.

Kane, R. 2005. *A Contemporary Introduction to Free Will*. Oxford University Press.

Kane, R. 2007. Libertarianism. In *Four Views on Free Will*, J. M. Fischer, R. Kane, D. Pereboom & M. Vargas (eds), pp. 5–43. Blackwell. www.thedivineconspiracy.org/Z5217X.pdf

Kant, I. 2002/1785. *Groundwork for the Metaphysics of Morals*. A. Zweig (trans.), T. E. Hill Jr. & A. Zweig (eds). Oxford University Press.

Kauppinen, A. 2007. The rise and fall of experimental philosophy. *Philosophical Explorations 10*(2): 95–118.

Kauppinen, A. 2014a. Empathy, emotion regulation, and moral judgment. In *Empathy and Morality*, H. L. Maibom (ed.), Oxford University Press.

Kauppinen, A. 2014b. Ethics and empirical psychology: Critical remarks to empirically informed ethics. In *Empirically Informed Ethics: Morality between Facts and Norms*, Christen, M. E., C. E. van Schaik, J. E. Fischer, M. E. Huppenbauer & C. Tanner (eds), pp. 279–305. Springer.

Keim, B. 2012. Brain scanners can see your decisions before you make them. *Wired*. www.wired.com/science/discoveries/news/2008/04/mind_decision

Kelly, D. R. 2011. *Yuck!: The Nature and Moral Significance of Disgust*. MIT Press.

Kennett, J. 2006. Do psychopaths really threaten moral rationalism? *Philosophical Explorations 9*(1): 69–82.

Keteyian, A. 2010. Madoff victims vent their anger in print. *CBS News*. March 18. www.cbsnews.com/8301–500690_162–5090670.html

Kitcher, P. 2011. *The Ethical Project*. Harvard University Press.

Kleingeld, P. 2014. Debunking confabulation: Emotions and the significance of empirical psychology for Kantian ethics. In *Kant on Emotion and Value*, pp. 146–165. Palgrave Macmillan.

Kohlberg, L., 1984. *Essays on Moral Development, vol. 2. The Psychology of Moral Development*. Harper & Row.

Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge University Press.

Kraut, R. 2009. *What Is Good and Why: The Ethics of Well-Being*. Harvard University Press.

Kumar, V., & R. Campbell. 2022. *A Better Ape: The Evolution of the Moral Mind and How It Made Us Human*. Oxford University Press.

Kurzban, R., M. N. Burton-Chellew, & S. A. West. 2015. The evolution of altruism in humans. *Annual Review of Psychology 66*(1): 575–599.

Landy, J. F., & G. P. Goodwin. 2015. Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science 10*(4): 518–536.

Latané, B., & J. M. Darley. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* Appleton-Century Crofts.

Latané, B., & S. Nida. 1981. Ten years of research on group size and helping. *Psychological Bulletin 89*(2): 308.

Lazarus, R. S. 1991. Cognition and motivation in emotion. *American Psychologist 46*(4): 352–367.

Lerner, J. S., J. H. Goldberg, & P. E. Tetlock. 1998. Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin 24*(6): 563–574.

Levy, N. 2015. Less blame, less crime? The practical implications of moral responsibility skepticism. *Journal of Practical Ethics 3*(2).

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences 8*(4): 529–566.

Little, M. O. 1997. Virtue as knowledge: Objections from the philosophy of mind. *Nous 31*(1): 59–79.

Lloyd, E. A. 1999. Evolutionary psychology: The burdens of proof. *Biology and Philosophy 14*: 211–233.

Lombrozo, T. 2009. The role of moral commitments in moral judgment. *Cognitive Science 33*(2): 273–286.

Lutz, Matthew, & James Lenman. (Spring 2021 edition). Moral naturalism. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/naturalism-moral/

Lyubomirsky, S. 2008. *The How of Happiness: A Scientific Approach to Getting the Life You Want*. Penguin.

Machery, E., & J. M. Doris. 2017. An open letter to our students: Doing interdisciplinary moral psychology. In *Moral Psychology*, B. Voyer & T. Tarantola (eds), Springer.

Machery, E., & R. Mallon. 2010. Evolution of morality. In *The Moral Psychology Handbook*, J. Doris and the Moral Psychology Research Group (eds), pp. 3–46. Oxford University Press.

Mackie, J. 1990. *Ethics: Inventing Right and Wrong*. Penguin.

Manne, K. 2014. Internalism about reasons: Sad but true? *Philosophical Studies 167*, 89–117.

Markey, P. M. 2000. Bystander intervention in computer-mediated communication. *Computers in Human Behavior 16*(2): 183–188.

May, J. 2014. Does disgust influence moral judgment? *Australasian Journal of Philosophy 92*(1): 125–141.

May, J. 2018. *Regard for Reason in the Moral Mind*. Oxford University Press.

McCann, K. 2012. Altruistic donation: "I could save a life – and that's all that matters." *The Telegraph*, November 22.

McDowell, J. 1979. Virtue and treason. *Monist 62*(3): 331–350.

Mele, A. R. 2006. *Free Will and Luck*. Oxford University Press.

Merritt, M. 2000. Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice 3*(4): 365–383.

Mesquita, B. 2022. *Between Us: How Cultures Create Emotions*. National Geographic Books.

Mikhail, J. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences 11*(4): 143–152.

Milgram, S. 1963. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology 67*(4): 371–378.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. HarperCollins.

Miller, C. 2003. Social psychology and virtue ethics. *The Journal of Ethics 7*(4): 365–392.

Miller, C. 2013. *Moral Character: An Empirical Theory*. Oxford University Press.

Miller, C. 2014. *Character and Moral Psychology*. Oxford University Press.

Murphy, M. C., & G. M. Walton. 2013. From prejudiced people to prejudiced places: A social-contextual approach to prejudice. In *Stereotyping and Prejudice*, C. Stangor & C. S. Crandall (eds), pp. 193–216. Psychology Press.

Nadelhoffer, T., & A. Monroe (eds). 2022. *Advances in Experimental Philosophy of Free Will and Responsibility*. Bloomsbury Publishing.

Nagel, T. 1970. *The Possibility of Altruism*. Clarendon Press.

Nahmias, E. 2011. Scientific challenges to free will. In *A Companion to the Philosophy of Action*, T. O'Connor & S. Constantine (eds), pp. 345–356. Wiley.

Nahmias, E. 2012a. *The psychology of free will. Oxford Handbook on the Philosophy of Psychology*. Oxford University Press. www2.gsu.edu/~phlean/papers/Nahmias_Psychology_of_Free_Will_prepublication.pdf

Nahmias, E. 2012b. Free will and responsibility. *Wiley Interdisciplinary Reviews: Cognitive Science 3*(4): 439–449.

Nahmias, E., S. G. Morris, T. Nadelhoffer, & J. Turner. 2007. Is incompatibilism intuitive? *Philosophy and Phenomenological Research 73*(1): 28–53.

Nelkin, D. K. 2008. Responsibility and rational abilities: Defending an asymmetrical view. *Pacific Philosophical Quarterly 89*(4): 497–515.

Nelson, S. K., K. Layous, S. W. Cole, & S. Lyubomirsky. 2016. Do unto others or treat yourself? The effects of prosocial and self-focused behavior on psychological flourishing. *Emotion 16*(6): 850.

Nichols, S. 2002a. How psychopaths threaten moral rationalism. *The Monist 85*(2): 285–303.

Nichols, S. 2002b. Norms with feeling: Towards a psychological account of moral judgment. *Cognition 84*: 221–236.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*, vol. *13*. Oxford University Press.

Nichols, S., & J. Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous 41*(4): 663–685.

Nisbett, R. E., & T. D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84*(3): 231–259.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Basic Books.

Nussbaum, M. 1985, January. Chapter 6: The discernment of perception: An Aristotelian conception of private and public rationality. In *Proceedings of the Boston Area Colloquium in Ancient Philosophy 1*(1): 151–201. Brill.

Nussbaum, M. C. 1995. Aristotle on human nature and the foundations of ethics. In *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*, J. E. J. Althan & R. Harrison (eds), pp. 86–131. Cambridge University Press.

Nussbaum, M. C. 2001. *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

Nussbaum, M. C. 2003. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press.

Nussbaum, M. C. 2009. *Hiding from Humanity: Disgust, Shame, and the Law*. Princeton University Press.

O'Connor, T. 2000. *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press.

O'Neill, E. 2017. Kinds of norms. *Philosophy Compass 12*(5): 1–15.

Pavot, W., & E. Diener. 2009. Review of the Satisfaction with Life scale. In *Assessing Well-being*, pp. 101–117. Springer.

Payne, B. K., H. A. Vuletich, & K. B. Lundberg. 2017. The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry 28*(4): 233–248.

Pereboom, D. 2001. *Living without Free Will*. Cambridge University Press.

Petrinovich, L., & P. O'Neill. 1996. Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology 17*(3): 145–171.

Piliavin, J. A. 2003. Doing well by doing good: Benefits for the benefactor. In *Flourishing: Positive Psychology and the Life Well-Lived*, C. L. M. Keyes & J. Haidt (eds), pp. 227–247. American Psychological Association.

Prinz, J. 2004a. Embodied emotions. In *Thinking about Feeling: Contemporary Philosophers on Emotions*, R. C. Solomon (ed.), pp. 44–58. Oxford University Press.

Prinz, J. 2004b. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press.

Prinz, J. 2006. The emotional basis of moral judgments. *Philosophical Explorations 9*(1): 29–43.

Prinz, J. 2007. *The Emotional Construction of Morals*. Oxford University Press.

Prinz, J. 2009. Against moral nativism. In *Stich and His Critics*, D. Murphy & M. Bishop (eds), pp. 167–189. Wiley.

Prinz, J. 2011. Against empathy. *The Southern Journal of Philosophy 49*, 214–233.

Prinz, J. 2012. *Beyond Human Nature: How Culture and Experience Shape Our Lives*. Penguin.

Raibley, J. 2010. Well-being and the priority of values. *Social Theory and Practice 36*(4): 593–620.

Railton, P. 1984. Alienation, consequentialism, and the demands of morality. *Philosophy and Public Affairs 13*(2): 134–171.

Railton, P. 1986. Moral realism. *The Philosophical Review 95*(2): 163–207.

Rawls, J. 1951. Outline of a decision procedure for ethics. *The Philosophical Review 60*(2): 177–197.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.

Rawls, J. 1980. Kantian constructivism in moral theory. *The Journal of Philosophy 77*(9): 515–572.

Rhodes, M., & H. Wellman. 2017. Moral learning as intuitive theory revision. *Cognition 167*: 191–200.

Richerson, P. J., & R. Boyd. 2008. *Not by Genes Alone: How Culture Transformed Human Evolution*. University of Chicago press.

Roberts, R. C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge University Press.

Rosati, C. S. 2009. Self-interest and self-sacrifice. *Proceedings of the Aristotelian Society 109*: 311–325.

Rosati, Connie S. 2016. Moral motivation. *The Stanford Encyclopedia of Philosophy* (Winter edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2016/entries/moral-motivation/

Rose, D., & S. Nichols. 2013. The lesson of bypassing. *Review of Philosophy and Psychology 4*(4): 599–619.

Roskies, A. 2003. Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy". *Philosophical Psychology 16*(1): 51–66.

Roskies, A. 2008. Neuroimaging and inferential distance. *Neuroethics 1*(1): 19–30.

Roskies, A. L. 2010a. How does neuroscience affect our conception of volition? *Annual Review of Neuroscience 33*, 109–130.

Roskies, A. 2010b. Why Libet's studies don't pose a threat to free will. In *Conscious Will and Responsibility: A Tribute to Benjamin Libet*, W. Sinnott-Armstrong & L. Nadel (eds), pp. 11–22. Oxford University Press.

Roskies, A. L. 2021. The neuroscience of free will. *U. St. Thomas JL & Pub. Pol'y 15*: 162.

Ross, L. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology 10*: 173–220.

Rottman, J., & D. Kelemen. 2012. Aliens behaving badly: Children's acquisition of novel purity-based morals. *Cognition 124*(3): 356–360.

Rottman, J., & L. Young. 2015. Mechanisms of moral development. *The Moral Brain: A Multidisciplinary Perspective*, pp. 123–142. MIT Press.

Ryan, R. M., & E. L. Deci. 2001. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology 52*(1): 141–166.

Ryff, C. D. 1989. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology 57*(6): 1069–1081.

Scanlon, T. M. 1982. Contractualism and utilitarianism. In *Utilitarianism and Beyond*, A. Sen & B. Williams (eds), pp. 103, 110. Cambridge University Press.

Scanlon, T. M. 1998. *What We Owe to Each Other*. The Belknap Press of Harvard University Press.

Scarantino, Andrea. 2014. The motivational theory of emotions. In *Moral Psychology and Human Agency*, Daniel Jacobson & Justin D'Arms (eds), pp. 156–185. Oxford University Press.

Schachter, S., & J. Singer. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychological Review 69*(5): 379–399.

Scherer, K. R. 2000. Psychological models of emotion. In *The Neuropsychology of Emotion*, Joan C. Borod (ed.), pp. 137–162. Oxford University Press.

Scherer, K. R. 2005. What are emotions? And how can they be measured? *Social Science Information 44*(4): 695–729.

Schlingloff, L., G. Csibra, & D. Tatone. 2020. Do 15–month-old infants prefer helpers? A replication of Hamlin et al. (2007). *Royal Society Open Science 7*(4), article 191795.

Schnall, S., J. Haidt, G. L. Clore, & A. H. Jordan. 2008. Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin 34*(8): 1096–1109.

Schroeder, M. A. 2007. *Slaves of the Passions*. Oxford University Press.

Schroeder, T. 2004. *Three Faces of Desire*. Oxford University Press.

Schroeder, T. 2006. Desire. *Philosophy Compass 1*(6): 631–639.

Schwitzgebel, E. 2011. Bartels and Pizarro: Consequentialists are psychopaths. *The Splintered Mind*. URL = http://schwitzsplinters.blogspot.com/2011/09/bartels-and-pizarro-consequentialists.html

Seidel, A., & J. Prinz. 2013. Sound morality: Irritating and icky noises amplify judgments in divergent moral domains. *Cognition 127*(1): 1–5.

Seligman, M. E. P. 2002. *Authentic Happiness: Using the New Positive Psychology to Realize Your Potential for Lasting Fulfillment*. Simon & Schuster.

Shafer-Landau, R. 2003. *Moral Realism: A Defence*. Oxford University Press.

Shafer-Landau, R. 2008. Defending ethical intuitionism. In *Moral Psychology*, vol. *2*, W. Sinnott-Armstrong (ed.), pp. 83–96. MIT Press.

Shariff, Azim F., & Ara Norenzayan. 2007. God is watching you. *Pyschological Science 18*(9): 803–809.

Shariff, Azim F., Aiyana K. Willard, Teresa Andersen, & Ara Norenzayan. 2016. Religious priming: A meta-analysis with a focus on prosociality. *Personality and Social Pyschology Review 20*(1): 27–48. doi:10.1177/1088868314568811

Shenhav, A., & J. Greene. 2014. Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience 34*(13): 4741–4749.

Shoemaker, D. 2015. *Responsibility from the Margins*. Oxford University Press.

Sifferd, K. L., & W. Hirstein. 2013. On the criminal culpability of successful and unsuccessful psychopaths. *Neuroethics 6*(1): 129–140.

Sin, N. L., & S. Lyubomirsky. 2009. Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: A practice-friendly meta-analysis. *Journal of Clinical Psychology 65*(5): 467–487.

Singer, P. 1972. Famine, affluence, and morality. *Philosophy & Public Affairs 1*(3): 229–243.

Sinnott-Armstrong, W. 2008. Framing moral intuitions. In *Moral Psychology, Vol. 2: The Cognitive Science of Morality: Intuitions and Diversity*, W. Sinott-Armstong (ed.), pp. 47–76. MIT Press.

Sinnott-Armstrong, W. (ed.). 2014. *Moral Psychology, Vol. 4: Freedom and Responsibility*. MIT Press.

Slater, L. 2004. *Opening Skinner's Box: Great Psychological Experiments of the Twentieth Century*. W. W. Norton.

Smetana, J. G. 1981. Preschool children's conceptions of moral and social rules. *Child Development 52*: 1333–1336.

Smetana, J. G. 1993. Understanding of social rules. In *The Development of Social Cognition: The Child as Psychologist*, M. Bennett (ed.), pp. 111–141. Guilford Press.

Smetana, J. G., M. Jambon, & C. Ball. 2013. The social domain approach to children's moral and social judgments. In *Handbook of Moral Development*, pp. 23–45. Psychology Press.

Smith, E. I. 2022. Thinking like a psychological scientist. In *Noba Textbook Series: Psychology*, R. Biswas-Diener & E. Diener (eds). DEF Publishers. URL = http://noba.to/nt3ysqcm

Smith, M. 1987. The Humean theory of motivation. *Mind 96*(381). New Series (January 1): 36–61.

Smith, M. 1995a. Internal reasons. Philosophy and Phenomenological. *Research 55*(1) 109–131.

Smith, M. 1995b. *The Moral Problem*. Blackwell.

Snow, N. E. 2010. *Virtue as Social Intelligence: An Empirically Grounded Theory*. Routledge.

Sober, E., & D. S. Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Solomon, R. C. 1973. Emotions and choice. *The Review of Metaphysics 27*(1): 20–41.

Sommers, T. 2005. Interview with Jonathan Haidt. *The Believer*. www.believermag.com/issues/200508/?read=interview_haidt

Soon, C. S., M. Brass, H. J. Heinze, & J. D. Haynes. 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience 11*(5): 543–545.

Sousa, R. De. 1979. The rationality of emotions. *Dialogue 18*(1): 41–63.

Sripada, C. 2016. Self-expression: A deep self theory of moral responsibility. *Philosophical Studies 173*(5): 1203–1232.

Stanley, M. L., S. Yin, & W. Sinnott-Armstrong. 2019. A reason-based explanation for moral dumbfounding. *Judgment and Decision Making 14*(2): 120–129.

Steinberg, D. 2003. Kidneys and the kindness of strangers. *Health Affairs 22*(4): 184–189.

Sterelny, K. 2010. Moral nativism: A sceptical response. *Mind & Language 25*(3): 279–297.

Stevenson, R. J., M. J. Oaten, T. I. Case, B. M. Repacholi, & P. Wagland. 2010. Children's response to adult disgust elicitors: Development and acquisition. *Developmental Psychology 46*(1): 165.

Stich, S. 2007. Evolution, altruism and cognitive architecture: A critique of Sober and Wilson's argument for psychological altruism. *Biology & Philosophy 22*(2): 267–281.

Stich, S., J. M. Doris, & E. Roedder. 2010. Altruism. In *The Moral Psychology Handbook*, J. Doris (ed.), pp. 147–206. Oxford University Press.

Stichter, M. 2007. Ethical expertise: The skill model of virtue. *Ethical Theory and Moral Practice 10*(2): 183–194.

Stichter, M. 2020. Learning from failure: Shame and emotion regulation in virtue as skill. *Ethical Theory and Moral Practice 23*(2): 341–354.

Stocker, M. 1976. The schizophrenia of modern ethical theories. *The Journal of Philosophy 73*(14): 453–466.

Stocks, E. L., D. A. Lishner, & S. K. Decker. 2009. Altruism or psychological escape: Why does empathy promote prosocial behavior? *European Journal of Social Psychology 39*(5): 649–665.

Stohr, K. 2022. *Choosing Freedom: A Kantian Guide to Life*. Oxford University Press.

Strawson, P. F. 2008. *Freedom and Resentment and Other Essays*. Routledge.

Street, S. 2010. What is constructivism in ethics and metaethics? *Philosophy Compass 5*(5): 363–384.

Sumner, L. 1996. *Welfare, Happiness, and Ethics*. Oxford University Press.

Swartwood, J. D. 2013. Wisdom as an expert skill. *Ethical Theory and Moral Practice 16*(3): 511–528.

Tappolet, Christine. 2016. *Emotions, Values, and Agency*. Oxford University Press.

Tappolet, Christine. 2020. Emotions inside out: The nonconceptual content of emotions. In *Concepts in Thought, Action, and Emotion: New Essays*, Christoph Demmerling and Dirk Schroeder (eds), pp. 257–276. Routledge.

Tappolet, Christine. 2023. *Philosophy of Emotion*. Routledge.

Tasimi, A. 2020. Connecting the dots on the origins of social knowledge. *Perspectives on Psychological Science 15*, 397–410.

Taurek, J. M. 1977. Should the numbers count? *Philosophy & Public Affairs 6*(4): 293–316. URL = www.jstor.org/stable/2264945. Accessed 12 Apr. 2022.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist 59*(2): 204–217.

Thomson, J. J. 2008. Turning the trolley. *Philosophy & Public Affairs 36*(4): 359–374.

Tiberius, V. 2008. *The Reflective Life: Living Wisely with Our Limits*. Oxford University Press.

Tiberius, V. 2009. The reflective life: Wisdom and happiness for real people. In *Philosophy and Happiness*, L. Bortolotti (ed.), pp. 215–232. Palgrave Macmillan.

Tiberius, V. 2013. In defense of reflection. *Philosophical Issues 23*(1): 223–243.

Tiberius, V. 2018. *Well-Being as Value Fulfillment: How We Can Help Each Other to Live Well*. Oxford University Press.

Tomasello, M. 2016. A natural history of human morality. In *A Natural History of Human Morality*. Harvard University Press.

Tversky, A., & D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science 211*(4481): 453–458.

Twain, M. 1994/1884. *Adventures of Huckleberry Finn*. Dover.

Van de Vondervoort, J. W., & J. K. Hamlin. 2017. Preschoolers' social and moral judgments of third-party helpers and hinderers align with infants' social evaluations. *Journal of Experimental Child Psychology 164*: 136–151.

Van Norden, B. (trans.) 2008. *Mengzi: With Selections from Traditional Commentaries*. Hackett.

Van Roojen, M. 2015. *Metaethics: A Contemporary Introduction*. Routledge.

Vargas, M. 2009. Revisionism about free will: A statement and defense. *Philosophical Studies 144*(1): 45–62.

Vargas, M. 2017. Implicit bias, responsibility and moral ecology. In *Oxford Studies in Agency and Responsibility*, D. Shoemaker (ed.). Oxford University Press.

Vargas, M., & J. M. Doris. 2022. *Oxford Handbook of Moral Psychology*. Oxford University Press.

Vihvelin, Kadri. 2022. Arguments for incompatibilism. *The Stanford Encyclopedia of Philosophy* (Fall edition), Edward N. Zalta & Uri Nodelman (eds). https://plato.stanford.edu/archives/fall2022/entries/incompatibilism-arguments/

Wallace, R. J. 1990. How to argue about practical reason. *Mind 99*: 355–385.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.

Washington, N., & D. Kelly. 2016. Who's responsible for this? Moral responsibility, externalism and knowledge about implicit bias. In *Implicit Bias and Philosophy* (Vol. *2*), M. Brownstein & J. Saul (eds), pp. 11–36. Oxford University Press.

Weinstein, N., & R. M. Ryan. 2010. When helping helps: Autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of Personality and Social Psychology 98*(2): 222–244.

Wheatley, T., & J. Haidt. 2005. Hypnotic disgust makes moral judgments more severe. *Psychological Science 16*(10): 780–784.

Wiggins, D. 1987. *Needs, Values, Truth: Essays in the Philosophy of Value*. Oxford University Press.

Williams, B. 1981. *Moral Luck: Philosophical Papers 1973–1980*. Cambridge University Press.

Wilson, E. O. 1975. *Sociobiology: The New Synthesis*. Harvard University Press.

Wilson, T. D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Harvard University Press.

Wilson, T. D., & D. Kraft. 1993. Why do I love Thee?: Effects of repeated introspections about a dating relationship on attitudes toward the relationship. *Personality and Social Psychology Bulletin 19*(4): 409–418.

Wilson, T. D., D. Kraft, & D. S. Dunn. 1989. The disruptive effects of explaining attitudes: The moderating effect of knowledge about the attitude object. *Journal of Experimental Social Psychology 25*(5): 379–400.

Wilson, T. D., D. J. Lisle, J. W. Schooler, S. D. Hodges, K. J. Klaaren, & S. J. LaFleur. 1993. Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin 19*: 331–331.

Wilson, T. D., & R. E. Nisbett. 1978. The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology*. http://psycnet.apa.org/psycinfo/1980–24471–001.

Wolf, S. 1980. Asymmetrical freedom. *The Journal of Philosophy 77*(3): 151–166.

Wolf, S. 1981. The importance of free will. *Mind 90*(359): 386–405.

Wolf, S. 1990. *Freedom within Reason*. Oxford University Press.

Woo, B., E. Tan, & J. Hamilin. 2022. Human morality is based on an early-emerging moral core. *Annual Review of Developmental Psychology 4*: 44–61.

Zagzebski, L. 2004. *Divine Motivation Theory*. Cambridge University Press.

Zhao, X., Liu, L., Zhang, X. X., Shi, J. X., & Huang, Z. W. 2014. The effect of belief in free will on prejudice. *PloS One 9*(3): e91572.

Index

Page references including ‘n’ denote endnotes

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com