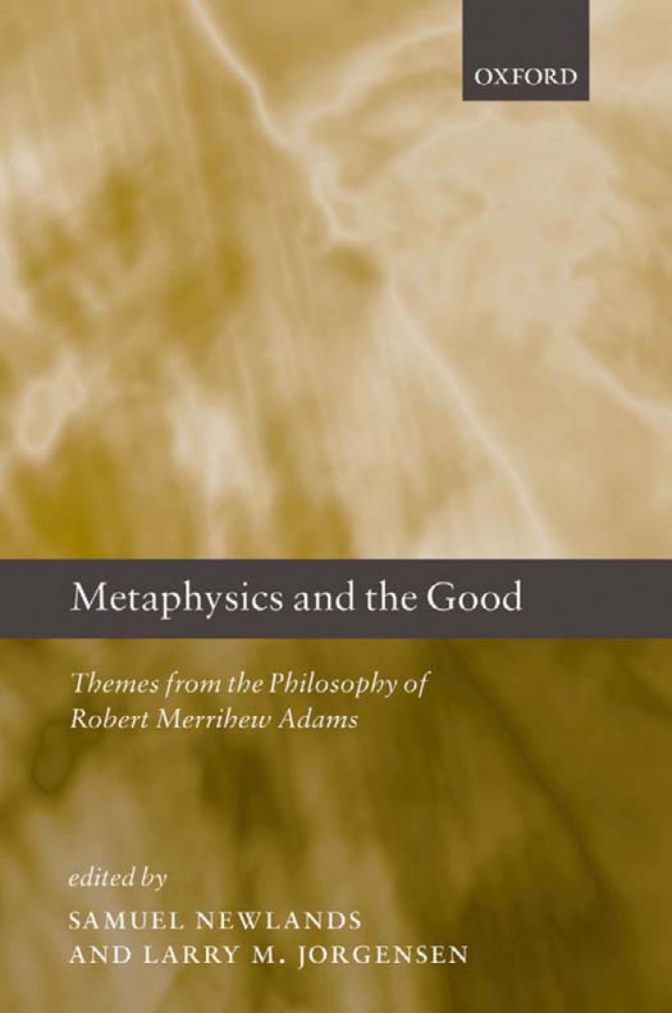
OXFORD
Metaphysics and the Good
Themes from the Philosophy of
Robert Merrihew Adams
edited by
SAMUEL NEWLANDS
AND LARRY M. JORGENSEN

Metaphysics and the Good

This page intentionally left blank

Metaphysics and the Good

Themes from the Philosophy of Robert Merrihew Adams

EDITED BY
Samuel Newlands and Larry M. Jorgensen

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2009

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Data available
Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire
ISBN 978–0–19–954268–0
10 9 8 7 6 5 4 3 2 1

For Bob

Acknowledgements

First and foremost, we would like to thank Robert Merrihew Adams. It borders on the trivial to suggest that this volume wouldn't have been possible without him. But, throughout his career, Bob has embodied an ethic of excellence, both in his philosophical work as well as in his kind and generous nature as a mentor to his students, and this has inspired us to put this volume together. The readiness with which our contributors participated in the project, the immediate and consistent encouragement of Oxford University Press, and the touching personal remarks made by everyone involved in the initial conference in honor of Bob all attest to his importance both to our profession and to his colleagues and students. We stand in a long line of those who have benefited personally and professionally from being Bob's students, and we intend this volume as a small representation of our gratitude to him.

We would like to extend a special note of gratitude to Michael Della Rocca, who has tirelessly helped and encouraged us with this project from the very beginning. We would also like to thank the following individuals for providing feedback and invaluable assistance along the way: Shelly Kagan, Michael Nelson, Sun-Joo Shin, and the participants and audience members in the 2005 conference 'Metaphysics, History, Ethics: A Conference in Honor of Robert Merrihew Adams'. We would also like to thank the contributors to this volume for their diligence and patience during the publication process. Peter Momtchiloff also deserves our gratitude for his general encouragement and assistance, as well as his help in deciding on a title for the volume. This project was made possible in part by support from the Institute for Scholarship in the Liberal Arts, College of Arts and Letters, University of Notre Dame.

Lastly, we would like to thank our families for their steadfast love, support, patience, and encouragement.

List of Contributors

Robert Merrihew Adams is a senior member of the Faculty of Philosophy at Oxford University.

Paul Hoffman is Professor of Philosophy at the University of California, Riverside.

Larry M. Jorgensen is Assistant Professor of Philosophy at Valparaiso University.

Shelly Kagan is Clark Professor of Philosophy at Yale University.

Michael Nelson is Assistant Professor of Philosophy at the University of California, Riverside.

Samuel Newlands is Assistant Professor of Philosophy at the University of Notre Dame.

Derk Pereboom is Professor of Philosophy at Cornell University.

Marleen Rozemond is Associate Professor of Philosophy at the University of Toronto.

R. C. Sleigh, Jr. is Professor Emeritus of Philosophy at the University of Massachusetts, Amherst.

Houston Smit is Associate Professor of Philosophy at the University of Arizona.

Jeffrey Stout is Professor of Religion at Princeton University.

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

Allen W. Wood is Ward W. and Priscilla B. Woods Professor of Philosophy at Stanford University.

Dean Zimmerman is Professor of Philosophy at Rutgers University.

This page intentionally left blank

Contents

This page intentionally left blank

Introduction

SAMUEL NEWLANDS AND LARRY M. JORGENSEN

When several of Robert Merrihew Adams's colleagues and students organized a conference at Yale University in honor of his retirement, we faced what proved to be a daunting question. What turn of phrase best encapsulates Adams's seminal work in so many different areas of philosophy—metaphysics, philosophy of religion, history of philosophy, and ethics? As inspiration proved elusive, despair set in. In the end, we settled for mere description and named the gathering 'Metaphysics, History, Ethics: A Conference in Honor of Robert Merrihew Adams'. Some of the papers in this volume were presented at that conference in the spring of 2005; others were solicited and added later. But all of the papers appear in print here for the first time, and all pursue the conference's original goal of honoring Adams by exploring and sometimes challenging the themes and topics that have animated his philosophical life.

But while we think the present volume's title is catchier, the more significant question behind our original naming dilemma remains. What thematically and systematically connects Adams's work on ontology, modality, identity, existence, idealism, arguments for the existence of God, the problem of evil, divine knowledge, faith, love, metaethics, virtue theory, divine-command theory, as well as on historical figures such as Leibniz, Descartes, Berkeley, Kant, Kierkegaard, and Schleiermacher? Having learned our lesson, we turned directly to the source this time and asked Adams what *he* thought tied together his philosophical interests and achievements. His reply constitutes the first essay in this volume, 'A Philosophical Autobiography' (Chapter 1).

In that essay, Adams is reluctant to enforce complete thematic unity on his own views. He does not, for instance, appeal to an original insight as

the wellspring for all of his subsequent philosophical projects. And rightly so—as he tells the story, his interest in some of the fields in which he made his most prominent contributions was kindled by very contingent factors, like teaching demands at early jobs. Nor does Adams provide a neat and comprehensive retrospective framework within which all his views neatly fit. This too seems right, especially given the ways his interests and contributions continue to evolve, expanding into his ongoing work on virtues, sexual ethics, and the nature of existence, to name a few.

The essays in this volume, summarized below, explore facets of many of Adams's conclusions in all of the most significant categories of his work. But, before turning to the essays themselves, we want to draw our readers' attention to several broader, interconnected themes that inform Adams's approach to philosophy itself. This is partly to summarize some of the more sweeping conclusions Adams makes in Chapter 1. But these broader themes also deeply informed our decisions about creating and editing this volume. We hope that by understanding a bit more about Adams's views on philosophy, readers will better understand what to expect in this collection.

In his autobiographical essay, Adams narrates his 'falling in love with philosophy', and this is no mere *façon de parler*. Love plays an important role in Adams's work in metaethics—for example, the loving character of God conditions what his divine-command theorist will accept as ethically binding. But the place of love in Adams's thought extends well beyond ethics and informs his view of the practice of philosophy itself. For Adams, the objects of philosophical reflection are objects accepting of and, more importantly, worthy of love. As she was for Plato, the philosopher for Adams is not paradigmatically a thinker, or a theoretician, or an experimenter, or an inventor, or even an admirer. She is a lover.

Certainly on some conceptions of philosophy, Adams's claim initially sounds incredible. If, for instance, one thought that the objects of philosophical study are primarily problems or confusions, such things may well seem hardly worthy of our love. Philosophy would be characterized by activities like solving and dissolving, not the pursuit and treasuring apropos to the lover. What is it about the objects of philosophical reflection that makes them even candidates for our love?

In another echo of Plato, Adams describes the realm of philosophy as 'full of objects of great beauty'. There is a beauty, Adams thinks, to the ways philosophers have raised, clarified, and engaged philosophical questions.

But though formal work in philosophy may invoke the sense of beauty akin to an elegant mathematical proof, it may be more difficult to see the beauty in, say, competing theories of direct reference or new interpretations of Aristotle's *De Anima*. So if it is the alleged beauty of philosophical objects that attracts our attention and affection qua philosophical lovers, what could explain the beauty of, for example, reflection dedicated to exhaustively and correctly completing the sentence, '*S knows that p iff* _____'?

Once again, Plato is not far from Adams's reply, as we ascend the ladder from beautiful things to beauty itself. The practice of philosophy is, Adams summarizes, a 'way of loving the Good'. The beauty of philosophical objects that attracts our love is grounded in their being reflections—some dimmer than others, no doubt—of the Good itself. So not only does philosophical activity turn out to be a way of pursuing beautiful objects, it is more fundamentally a loving that is directed, perhaps unaware, at the ultimate source of such beauty and goodness.

However sympathetic one may be towards these Platonic metaphors, it may still be easy to lose a grip on how such lofty imagery connects up to the actual and unwieldy collection of work that the profession has churned out over the last half-century and to which Adams himself has contributed significantly. Here it will be helpful to shift from Plato to another of Adams's philosophical soulmates: Leibniz. Leibniz famously believed that modal truths, propositions about the way the world actually is, might have been, or must be, are ultimately grounded in God's intellect. Surveying the structure and content of possible worlds is surveying, in a sense, divine real estate. In fact, Leibniz himself thought that one could validly reason from the existence of necessary truths to the existence of God. Whether or not he thinks such an argument ought to *persuade* anyone, Adams is deeply attracted to this broadly Leibnizian picture, according to which accounts of the ways the world might be are in fact tracing structures grounded in the Divine mind. Add in the further claim that philosophical theories are attempts to articulate the ways the world is, or might have been, or must be, and we begin to understand why Adams would describe philosophical activity as a way of loving the Good and the Beautiful. Philosophical reflection, at bottom, is a form of religious devotion. (As Adams emphasizes, of course, philosophy need not proceed or even understand itself explicitly within any such religious framework to be successful. Philosophy is also 'worth loving for its own sake', he affirms.)

It is here, hovering between Plato and Leibniz, that many of Adams's most pervasive philosophical commitments also coalesce: Christian theism, Neoplatonism, moral realism, and metaphysical idealism.

While not outright skeptical,[1] Adams doubts that philosophy has succeeded in definitively answering many of the questions that it has posed to itself about the contours of the world. The perennial openness of many of philosophy's central questions surely supports this doubt. On the other hand, Adams thinks philosophers have made significant progress in what he describes as 'exploring possible ways of thinking, giving us a clearer, deeper, and fuller understanding of them'. In fact, Adams adds that the flourishing of analytic philosophy along this dimension may be unmatched since the high Scholastic period in the thirteenth and fourteenth centuries. He notes, however, that such flourishing usually comes when we are engaged in systematic inquiries—or inquiries about the systematic inquiries of others. 'Philosophy resists piecemeal treatment . . . Philosophical theses tend to be fragments of actual or potential systems, and to look quite different in different systematic contexts'. This becomes most clear when we critically engage the systematic thoughts of others, an activity Adams likens to 'pulling a string here to see what moves over there, so to speak'.

Adams's own work on the history of philosophy exemplifies this approach. When he engages Leibniz, it is neither as a deferential apologist nor as an unsympathetic critic. He prods Leibniz's views and tries to stretch them, sometimes in new and uncomfortable directions, but always with the hope of shedding fresh light back on Leibniz's original views. It is through such experimental poking that we gain a greater understanding of not only Leibniz but also the ways in which philosophical concerns and questions hang—or fall—together. This kind of interpretive engagement with the systems of others reminds us, Adams writes, of 'the intrinsic systematicity of the subject matter, the interrelatedness of the problems of philosophy'.

And although Adams's work on figures in modern philosophy shows his giftedness for such philosophical interpretation, his work also contains some system-building of its own. And so, in this volume, we propose to

[1] Though Adams has identified himself as a 'skeptical realist' in philosophical theology (Robert M. Adams, *The Virtue of Faith and Other Essays in Philosophical Theology* (New York: Oxford University Press, 1987), 5).

explore Adams's systematic views in the very way in which he commends us to approach the systems of others—namely, with a good deal of poking and prodding (metaphorically speaking, of course). Instead of merely summarizing or reflecting on Adams's views, contributors were asked to honor Adams by engaging his views in this more exploratory way. Some take as a starting-point a conclusion of Adams's and run with it in new directions. Others attempt to replace a key idea of Adams's and explore the consequences of such a modification. Yet others attempt to insert Adams's views into a new context of discussion to see what light may be shed on both the context and Adams's original views. In all cases, we believe that the benefits of what Adams himself has done to others are earned here as well. Not only do we gain a greater understanding of possible ways of thinking about a range of philosophical topics, we also gain a greater understanding of the contours of and connections within Adams's thought. And that, we believe, is both a rewarding and intrinsically excellent advance.

Dean Zimmerman begins this exploration in his essay 'Yet Another Anti-Molinist Argument' (Chapter 2) by continuing one of Adams's own philosophical projects. To set the context, we begin with a small historical note: although mainstream philosophy may be, in at least some quarters, growing friendlier to theistic belief, Adams's work in the 1960s and 70s occurred in an environment in which philosophical theology was regarded with considerable suspicion if not outright hostility. Indeed, it was the work of well-trained analytic philosophers like Adams that helped earn philosophy of religion at least a grudging respectability. This rebirth of philosophy of religion in Anglo-American philosophy witnessed, among other things, a rebirth of one of the more hotly debated controversies in sixteenth-century philosophical theology. The controversy surrounded the range of God's knowledge. In the sixteenth-century version, which simmered down only after Papal intervention, almost everyone agreed that propositions describing future free actions of creatures have a truth-value that is known by God. (One recent development in analytic philosophy of religion has been the emergence of a wide range of voices challenging this point of agreement.) The controversy concerned propositions stating what creatures *would* freely do independently of and logically prior to God's creation of the actual world (now often called 'counterfactuals of freedom' (henceforth CFs)). So, if I had slept in yesterday, would I have freely eaten

breakfast at noon? As Adams colorfully replied, 'God only knows.' *Dramatic pause*. 'Or does He?'[2]

Adams has led the charge against those who would answer affirmatively, those known as 'Molinists' (so named after the sixteenth-century Jesuit Luis de Molina). The great allure of Molinism is a way of reconciling God's risk-free, sovereign control over the world with creaturely libertarian freedom. Adams and others have argued repeatedly that this attempted reconciliation fails. As part of his most influential criticism of Molinism, Adams argues that the truth-values of CFs would have to be groundless or *brutely* true or false. In reply, modern-day Molinists have generally accepted the brute nature of CFs and have focused on showing that this admission does not entail any untoward consequences. In his essay, Zimmerman counters that the Molinist concession to Adams and others about the bruteness of CFs forces Molinists to concede also a surprising possibility: it is possible that an omnipotent God couldn't have created free creatures in the first place. There are possible worlds in which the CFs line up in such a way that God would have freedom-undermining manipulative control over the outcome of such worlds. Molinists should admit, in other words, that it is metaphysically possible that the proposed Molinist reconciliation of divine sovereignty and human freedom fails. And, Zimmerman argues, admitting that *this* is a metaphysical possibility is unacceptable for both Molinists and non-Molinists alike.

In his 'The Contingency of Existence' (Chapter 3) Michael Nelson begins by considering the technical and more broadly philosophical problems of rejecting necessitarianism, the thesis that the actual world is the only possible world. Despite the fact that most of us have strong intuitions that the world might have gone differently, it is notoriously hard to cash out in both formal and metaphysical ways the claims that the actual world could have contained fewer or more objects than it actually does. After establishing both the formal and metaphysical pressures towards necessitarianism, Nelson discusses a number of different disarming strategies aimed at preserving our intuitions of contingent existence. As Nelson describes the most plausible responses, all claim that the apparent threat of necessitarianism rests on an equivocation; different respondents disagree on what they diagnose as the fundamental confusion. Does the threat of necessitarianism

[2] Adams, 'Middle Knowledge and the Problem of Evil', ibid., 77.

rest on a confusion between (a) modally discriminating and modally promiscuous natures (anti-essentialism); (b) ways of being versus ways of existing (Meinongianism); (c) contingent concreteness and contingent existence (Linsky and Zalta); (d) actually existing objects and non-actually existing objects (Lewisian possibilism); (e) unexemplified and exemplified Platonic essences (Plantinga); or (f) the way things are *in* a possible world and the way things are *at* a possible world (Adams, et al.)? Nelson argues that although each of these options has more plausibility than their critics, including Adams, admit, all but one are burdened with unnecessarily high theoretical and ontological costs. The winner, Nelson argues, is a version of Adams's own Aristotelian actualist solution, which has the further advantage of motivating a plausible solution to the technical concerns of non-necessitarianism as well.

In 'Consciousness and Introspective Inaccuracy' (Chapter 4) Derk Pereboom offers a broadly Kantian response to Frank Jackson's knowledge argument, and he argues that such a response can mitigate the challenge the knowledge argument poses to physicalism. Adams, in his 'Flavors, Colors, and God',[3] has argued that there is an explanatory gap between physical and phenomenal properties, and the prospects of closing the gap are very slim. If there is a distinction between the phenomenal qualities, like the sensation of red, and physical qualities, we can always raise the question why these phenomenal qualities are correlated with the particular underlying physical qualities. Why aren't the physical states correlated with a *different* phenomenal property, or none at all? In response to this challenge, Pereboom argues that there is an unexplored open possibility—namely, that introspection is inaccurate, that certain phenomenal properties are not represented accurately via introspection. If this is true, then there may be no gap between the *accurate* representation of phenomenal properties and the physical properties with which they are correlated.

Pereboom argues that a causal account of introspective representation would support this thesis. In a way analogous to sensory representation, where our knowledge of external objects is mediated by sensory representations that are caused by them, Pereboom argues (along Kantian lines) that the introspective representation of phenomenal properties is similarly mediated. Given this mediation, it is an open possibility that the introspective

[3] Adams, 'Flavors, Colors, and God', ibid., 243–62.

representations represent phenomenal properties inaccurately (i.e., 'as having a qualitative nature that they really lack'). This epistemic possibility allows Pereboom to give the following response to the knowledge argument: Mary's complete physical knowledge provides her with all she needs to represent accurately the real nature of the new phenomenal state that she encounters on leaving the room (i.e., her representation of a tomato as red). The phenomenal property represented introspectively as seeing red may not be as it is introspectively represented. So, on Pereboom's open possibility the following disjunction would apply to Mary: either (a) there is no phenomenal property represented by Mary's introspective representation of red, and so all Mary acquires in leaving the room is a false belief, the belief that there is such a phenomenal property, and so she comes to know nothing new, or (b) there *is* a phenomenal property represented by Mary's introspective representation of red, but that property is not *accurately* represented; the full and accurate representation is included in what Mary already knew while in the room, and so again she comes to know nothing new.

Whereas Pereboom has offered a broadly Kantian account of introspection, Houston Smit explores the concept of the a priori in Kant in 'Kant on Apriority and the Sponteneity of Cognition' (Chapter 5). In his book on Leibniz, Adams points out that the notions of a priori and a posteriori proof underwent a transition in the early modern period. Originally, 'a priori' and 'a posteriori' proofs were proofs 'from the cause' and 'from the effects', respectively,[4] and it is this sense that can still be found in the *Port Royal Logic*, for example. But at some point in the early modern period a transition was made from these notions of a priori and a posteriori proof to a new sense in which 'a priori' simply meant 'non-empirical'. Adams goes on to suggest that 'Leibniz played a crucial role in the transformation of the meaning of "a priori".'[5]

However, Smit argues that the earlier notion of 'a priori', what he calls the 'from-grounds' notion, can be found in Kant as well, and that the transition from this notion of a priority to the newer sense is a consequence of Kant's system. Smit offers substantial evidence in favor of his thesis that Kant was operating with the 'from-grounds' notion of the a priori,

[4] Adams, *Leibniz: Determinist, Theist, Idealist* (New York: Oxford University Press, 1994), 109–10.
[5] Ibid.

which he thinks clarifies Kant's project in the *Critique of Pure Reason*. In addition to this, Smit's argument provides a nice way of sharpening the differences between Leibniz and Kant. If both are operating from similar notions of the a priori—namely, the 'from-grounds' notion—how is it that they come to rather different conclusions? Smit argues that, given Kant's account of cognition, the 'non-empirical' notion of the a priori is entailed by the 'from-grounds' notion of the a priori, though there are no concepts of singular things by which one (perhaps with a sufficiently expansive mind) could come to know singular things a priori. That is, unlike for Leibniz, the a priori for Kant does not give us any grip on the things-in-themselves. Smit argues that Kant's account of the genesis of a priori thought is illuminated by the recognition that he is working not with a simply 'non-empirical' notion of the a priori, but with the 'from-grounds' notion of the a priori.

Robert Sleigh joins Adams in challenging recent scholarship on Leibniz's account of moral necessity in his essay 'Moral Necessity in Leibniz's Account of Human Freedom' (Chapter 6). Some have argued that Leibniz is operating with a concept of moral necessity derived from that found in earlier Spanish Jesuit philosophies. According to the Jesuit theory of moral necessity, Sleigh says, whenever a human agent makes a choice, this choice is *morally*, but not *metaphysically* necessary. This means, roughly, that the agent has the power to choose otherwise, that making a different choice in the same circumstances would not be contradictory nor miraculous, and that God nevertheless knows infallibly what the agent would choose were the circumstances to obtain.

Sleigh argues that this view is inconsistent with Leibniz's view that all properties of a substance are intrinsic to that substance (the 'doctrine of superintrinsicalness') and that all future states are caused by the prior internal states of the substance alone (the 'principle of spontaneity'). Similarly, the principle of spontaneity works against this theory of moral necessity, Sleigh says. The moral necessitarians used the locution 'inclines without necessitating' to mark those free actions that rational agents make, and which would not apply to non-rational agents. Sleigh argues instead that Leibniz intends the locution to apply equally to non-rational agents. According to Sleigh, Leibniz means simply to say that when some particular state of affairs obtains, it is *causally*, but not *metaphysically*, necessary that the effect obtains as well (causation understood relative to the appropriate domain). And, given the principle of spontaneity, *all* effects are causally

necessitated by a being's prior states alone (miracles aside), and so, for Leibniz, moral necessity extends much further than the earlier moral necessitarians intended.

In 'Leibniz on Final Causation' (Chapter 7) Marleen Rozemond develops an account of Leibniz's theory of causation that differs from the one Sleigh offers at the end of his essay. Rozemond argues that despite Leibniz's apparent separation of efficient and final causation to two separate domains (efficient causation operating at the level of bodies, final causation operating at the level of monads), Leibniz in fact allowed *both* efficient and final causation at the level of monads. This is in the spirit of an argument of Adams's that what is denied at the monadic level is not efficient causation, but merely *mechanical* causation. Rozemond situates Leibniz's theory of causation against the background of scholastic theories of causation, a necessary step if one is to understand Leibniz's revival of substantial forms and the final causes that are unique to them.

One objection to final causation that was prominent even among certain scholastics is that final causation requires knowledge of the end, and so final causation requires a mental substance. By reviving the notion of substantial forms, Leibniz at the same time revived the possibility of final causation. However, given the requirement of mentality for final causation, it could operate only on the level of simple substances. And so, given that bodies are not simple, Leibniz separates the domains of final and efficient causation, a separation that would have been foreign to the Aristotelians. But Rozemond goes on to argue that the separation is not complete. She identifies texts that suggest a sort of efficient causation even at the monadic level. What distinguishes the two realms is not an *absence* of efficient causation at the level of monads, rather it is the *presence* of final causes. But if this is so, how do we make sense of Leibniz's frequent claims about the divisions of the realms of final and efficient causation? Rozemond believes that beyond the basic causal story, the two realms are more *intelligible* through their respective forms of causation—bodily motion is made intelligible through efficient mechanical causes, and mental activity is made intelligible through intentions and volitions, the *ends* towards which the activity is directed.

In his 'Does Efficient Causation Presuppose Final Causation?: Aquinas vs. Early Modern Mechanism' (Chapter 8) Paul Hoffman pursues the same issues as Rozemond, but moves in a very different direction. Hoffman

answers the question in his title in the affirmative: efficient causation *does* presuppose final causation. So, even mechanical causation will involve final causes. In his argument, Hoffman appeals to Aquinas' account of final causation to show that the early modern philosophers were wrong to think their theories of motion had dispensed with final causes.

According to Hoffman, Aquinas has a core notion of final causation, as a cause tending to some end, and a full-bodied notion of final causation, which adds additional requirements, such as the requirements that the end be a good and that the agent acts with the purpose of achieving that end. The core notion of final causation is presupposed by efficient causation, Hoffman says, since any cause is directed at some particular effect—if unhindered, the cause will bring about the specified effect. This meets the minimal requirements of Aquinas' core notion of final causation—namely, that the cause tends to some particular effect rather than another. The early modern philosophers resisted the use of final causation in natural philosophy, and they provided an account of inertia that was supposed to undermine the teleological explanation of nature. Hoffman considers the views of Descartes, Newton, and Spinoza and argues that their theories do not really dispense with final causes. Hoffman adds that the stripped-down version of final causation is robust enough to remain philosophically interesting.

In his 'Herder and Kant on History: Their Enlightenment Faith' (Chapter 9) Allen Wood corrects what he sees as a gross misunderstanding of the relation between Kant and Herder on the nature of human history. In the course of developing this interpretive point, Wood also provides a broader defense of an Enlightenment philosophy of history against its more recent critics. Such critics charge that the horrors of the twentieth century alone put the lie to any 'naive' Enlightenment belief in an objective purpose that guides human history towards some grand, telic realization of reason and rationality. The spread of the ideals of autonomy, freedom, and rationality, it is charged, have obviously and dramatically failed to usher in a new golden age of peace and justice. In this polemic, Herder is sometimes invoked as an important counter-Enlightenment voice who stood against any such 'naive' Kantian faith in the ideal of historical progressivism.

All of this—the charge of naivety, the rejection of an intrinsic, unfolding purpose to human history, and the appeal to Herder by

Enlightenment critics—is fundamentally wrong, according to Wood. After sorting out the points of disagreement and, far more importantly, the broadly Enlightenment framework of agreement between Herder and Kant, Wood appeals to a secularized version of Adams's 'moral faith' to rebut the charge that Enlightenment approaches to human history are naive, immune to revision, or subject to dangerous totalizing tendencies. Such an orienting commitment to historical purpose, Wood uses Adams to draw out, is compatible with a healthy skepticism about humanity's realization of Enlightenment ideals, a keen awareness of our widespread and horrific failures to date, and a renewed effort to pattern our understanding of the direction of history after such revisable ideals. Indeed, Wood presses, our own need to exercise moral agency in response to such horrors renders such ideals necessary.

Susan Wolf also experiments with a secularized version of one of Adams's conclusions, focusing on his work on moral obligation in her 'Moral Obligations and Social Commands' (Chapter 10). Adams has proposed that our moral obligations are grounded in the commands of a loving God, arguing that such a divine-based theory of obligation, one whose content is often revealed and embedded in our social obligations, provides moral demands with their requisite objectivity and determinateness. Wolf is joined by Jeffrey Stout ('Adams on the Nature of Obligation' (Chapter 11)) in exploring whether Adams's attempt to explain moral obligations in terms of social obligations can be had without the high costs sometimes associated with additional appeals to God.

Stout's essay begins by worrying that Adams is inconsistent in his efforts to determine what the commands of the loving God are in the first place. Stout points out that Adams's Christian scriptures contain portrayals of divine commands to engage in activities that Adams himself admits are morally impermissible, such as murder and genocide. So, applying Plato's *Euthyphro* dilemma, either Adams's theory admits morally impermissible actions as permissible, or else there exists some further side-constraint on what counts as a revealed and morally binding command of God. Stout thinks that Adams accepts the second horn by limiting what counts as a revealed and binding command of a loving God based on its coherence with either our moral intuitions about goodness or with our existing social practices. But then it begins to sound like these moral intuitions and actual social practices are really doing the explanatory work

in accounting for our moral obligation in the first place. At the very least, Stout argues, we seem no better off at discerning our de facto moral obligations for having adopted a metaphysically loaded divine-command theory than if we had simply tried to read off our obligations by examining actual societal practices and intuitions without all the metaphysical baggage.

Adams might retort that such a secularized version of his social-command theory fails to provide the metaphysical objectivity and determinateness that moral obligations are supposed to have. However, Stout replies, once we turn to examining actual social practices as the source for trying to read off our de facto moral obligations, we lose confidence that our obligations are or must be so determinate in the first place. And so not only do the metaphysical posits of Adams's divine-command theory create difficult epistemic burdens, the best way of discharging them may call into question the very motivation for positing the theistic framework in the first place.

Wolf agrees with Stout that Adams's appeals to divine commands in this context are both unnecessary and problematic. She then uses several of Adams's insights to develop an account of moral obligations that is based on our actual social conventions, *sans* Adams's theological appeals. Wolf first argues in support of Adams that moral obligations are neither extensionally nor intensionally identical to the dictates of reason and offers several arguments to support the need for another source. But she thinks that a suitably developed theory of social obligations can do just that. Of course, as Wolf admits, there will be associated epistemic burdens with a secular social command theory, such as the difficulties in spelling out the identity conditions for societies and societal membership. But, even more threatening to the theory, isn't history populated with examples of societies that have commanded its members to behave in ways that we take to be morally impermissible? Must we now swallow the *first* horn of the *Euthyphro* dilemma and admit that our moral obligations are so plastic and historically relativized that even genocide might be morally permissible for some? Instead of appealing to the good and loving character of God, Wolf's proposed constraint appeals back to the deliverances of good moral reasoning as a necessary, but in itself insufficient basis for moral obligation. She notes in conclusion that her proposal may have the effect of diminishing the importance of the category of moral

obligation altogether. But, she argues, such a loss would actually be morally beneficial.

Shelly Kagan's essay 'The Grasshopper, Aristotle, Bob Adams, and Me' (Chapter 12) wonders what we would do in a Utopia. Supposing, he says, that all technological limitations have been overcome, there is no scarcity of any kind, and personal conflict has been eliminated. What would we do? Kagan points out that it is initially tempting to view this thought experiment as a way to get at what is intrinsically valuable, but, he argues, this would be a mistake. There are, he says, intrinsically valuable instrumental values. But, drawing on suggestions made by Adams, Kagan claims that perhaps no society, not even a utopian society, could contain all intrinsically valuable activities. But this should not deter us from asking the question, since in the end a Utopia may be a preferable society overall, even if it comes at the cost of some intrinsic goods.

So, returning to the original question, what would we do in Utopia? Kagan considers Bernard Suits's suggestion that utopian activity will be limited to game-playing and argues that this does not provide a sufficiently rich life to be a desirable one. Here again, Kagan develops an idea Adams discusses—the suggestion that well-being consists in the enjoyment of the excellent, which Kagan glosses as a pleasure in the possession and consumption of intrinsic goods. This will include such activities as the contemplation of the nature and laws of the universe, the contemplation of God, and the appreciation of beauty. This helps us appreciate the possibility of certain relations that cannot be separated from the activities of relating—to *be* in a certain relation with someone entails that we engage in *the activity* of relating. This, Kagan argues, doesn't appear to be an artificial constraint, and so not a game, and yet it does appear to add a good to one's life. So, one thing we will do in Utopia is to relate to one another—this will be an activity worth doing for its own sake.

This is the sort of activity we now turn over to the reader. The essays in this volume contribute to the common projects of exploring themes in Adams's work and advancing ideas that might be suitably applied in other philosophical systems. We believe that in doing so they shed new light on both Adams's own views and abiding questions in metaphysics, philosophy of religion, history of philosophy, and ethics. Can Adams's positions handle these challenges, proposed modifications, or new applications? Or would the philosophical theories on display in

this volume be more fruitfully embedded in alternative ways of trying to understand the world? These are questions we hope this volume stimulates among our readers. For Adams, reflecting on and wrestling with such questions is a way of loving the good. For us, it's also a way of honoring Adams.

1

A Philosophical Autobiography

ROBERT MERRIHEW ADAMS

My colleagues at Yale, generously organizing a conference in my honor, asked me to give an address at the banquet. I found it difficult to decide what to talk about. The idea came to me only the day before: when, if not on such an occasion, would it be appropriate for me to indulge in autobiographical reflection in public? Writing up my remarks in the present essay, I am rethinking as well as reconstructing them from the page of notes I had written down before the talk.

Is there a unity to my philosophical concerns? Their diversity made it hard to find a thematically unified title for the conference. To me, however, they seem to hang together; and one of my aims in narrating my life as a philosopher is to trace ways in which they have become integrated over the years. Still, I don't want to impose too complete a unity on them, either in narrative or in life. Each philosophical question demands attention in its own terms; and if one goes on learning, integration of one's views is a never-ending task.

I

I begin the story with my earliest memory of engaging in philosophical reasoning of any consequence. When I was fourteen or fifteen I became an idealist of a Berkeleyan sort. I had not heard of Berkeley, but I had recently been taught the modern, subjectivist view of colors, tastes, smells, and other so-called 'secondary qualities', which forms a starting-point for idealist argument in his *Three Dialogues*. I remember sitting on the lawn on a bright summer day, and wondering what a blade of grass could be like in itself. What could be its intrinsic qualities if the vivid green color and

the fresh grass smell were merely aspects of the way the grass affected my senses? The size and shape of the blade of grass were still supposed to be 'primary qualities', and real enough. But I was left with the question what it was that existed inside the space defined by those geometrical properties. What could it be like, in itself, for grass to exist in that space, rather than something else or nothing at all? I couldn't imagine what qualities could fill the space with reality if colors, tastes, and smells were ruled out as subjective.

Such questions led me to idealist thoughts. Should I really believe there is anything the grass is 'like' in itself? Maybe its reality is located where the vivid qualities are. Perhaps, that is, it exists only in my seeing, feeling, and smelling. I won't claim that I worked out a complete idealist theory. But I remember that I did ask myself why different people have similar perceptions (as I unskeptically assumed they do) if what we perceive has its reality in our personal perceptions. And I gave myself the same theological answer that Berkeley had given.

A year or so later, in reading, I encountered Berkeley's name, and the formula that 'to be is to be perceived'. I was ready to call myself a Berkeleyan. I wouldn't say exactly that about myself now. I suspect that Leibniz, in his panpsychist version of the ontological primacy of the mental or quasi-mental, may have been closer to the truth than Berkeley. But I continue to have broadly idealist views. I still doubt that any wholly unperceiving thing could exist as a thing in itself.

But that is not the main thread through my philosophical biography. A more organizing theme can be found in the reading I was doing when I finally met Berkeley's name and the words *esse est percipi*. It was in a book of theology by Paul Tillich. My later teenage years and my early twenties were a time of both deeper appropriation of Christian faith and intense wrestling with religious doubts and puzzlements. I was driven to theology, and eventually to philosophy, by a religious need to think through for myself questions about God and about Christianity. When it came time to choose my undergraduate major, I seriously considered history and classics as well as philosophy. I didn't think I would enjoy history or classics less; I chose philosophy because it seemed more important or more urgent. At a personal level it was what I needed to think about. As a matter of religious vocation also I had decided before going to college that I should become a minister; by the time I chose my major I had come to think it would be part

of my vocation to be a theologian. And it seemed to me that philosophy was the most important intellectual discipline for theology. I still hold that view about the relation of philosophy and theology, unfashionable as it may have become in theology.

II

Philosophically I was fortunate to enter Princeton University as an undergraduate in 1955, the year in which Gregory Vlastos and Carl Hempel arrived to play their central part in building the great philosophy department that Princeton has had for many decades now. The two philosophy courses I took in my first year were historical, but by the end of the year I had begun to be clued in to analytical philosophy. I bought A. J. Ayer's *Language, Truth, and Logic*, and read it during the following summer. I was immensely impressed by it. When I first had to face a class as a teaching fellow, several years later at Cornell, I was surprised at how difficult it was to explain why I had ever thought the verifiability criterion of meaning plausible enough to be worth worrying about. But in the summer after my freshman year at Princeton, I was almost persuaded of the fundamental soundness of Ayer's version of logical empiricism. And for several years I saw myself as thinking about philosophy, including the philosophy of religion, in an empiricist framework.

Among several outstanding teachers at Princeton, the one who did the most to excite and deepen my interest in analytical philosophy was Hilary Putnam, then an assistant professor there. I think the best philosophy course I ever took was a course in 'Advanced Logic' that Putnam co-taught with Paul Benacerraf, then still a graduate student. Their lectures covered a lot of logical theory, concluding with a fairly full sketch of Gödel's proof of his famous incompleteness theorem, and a discussion of its implications. That was excellent, but even more important for me were the 'preceptorials' (discussion sections), which I had with Putnam. Each week we read and discussed one of the great papers in philosophical logic from the previous six decades or so, including: Russell, 'On Denoting'; Frege, 'On Sense and Reference' (on *Sinn* and *Bedeutung*); Tarski on truth; Carnap, 'Empiricism, Semantics and Ontology'; Quine, 'On What There Is' and 'Two Dogmas of Empiricism'; and others. From my years at Princeton, and especially

from Putnam and Hempel, I retain a conception of analytical philosophy that owes more to its German than its British roots, and was shaped by interests in logic and philosophy of science.

At Princeton in the late 1950s all undergraduates wrote two junior essays that were term-long projects, and a senior thesis that was a year-long project. My junior essays were both historical, on Kant and Aristotle. Vlastos thought well enough of the Aristotle essay that he offered to advise me in the project if I wanted to develop it further as a senior thesis. I would have been wise to accept the offer. And I might have been even wiser to expand my essay on Kant's argument for the causal principle into a senior thesis. I think it was the most interesting thing I wrote as a student, graduate or undergraduate. I was essentially self-taught on Kant. I took on the project because I saw Kant as a philosopher I really needed to understand, and my introduction to him, in a Descartes to Kant course, had been by way of his *Prolegomena*, which has always seemed to me to leave out too much of what is most interesting and illuminating in the critical philosophy. I worked enormously hard on the argument about causality in Kant's first *Critique*, and came up with an interpretation similar to those that Peter Strawson and Jonathan Bennett were soon to publish, though of course much less fully developed than theirs.

For my senior thesis, however, my sense of my vocation led me to choose the topic of the use of language in prayer. The result was a disappointment to me, and I suspect to my advisers, Hugo Bedau and Sylvain Bromberger. It was an occasion for beginning to learn that in choosing a topic for philosophical work, the importance of the topic can matter less than the likelihood that one will have something to say that makes a difference to the discussion of the topic. It took me a long time to learn the lesson; and I fear I have remained subject to temptation in this area. Of course, it is also permanently difficult to discern what one will have something worth saying about.

III

The six years that followed my graduation from Princeton University in 1959 were devoted first to the study of theology for two years at Oxford and one year at Princeton Theological Seminary, and then to the practice of

ministry for three years as pastor of a small Presbyterian church in Montauk at the eastern tip of Long Island. During that whole time I continued to study philosophy as well as theology.

My theological program at Oxford was demanding, and I attended few classes in the philosophy faculty that were not about philosophy of religion. I went to all of J. L. Austin's 'informal instructions' in the last term he taught before his untimely death, and was awed by the performance, though I'm not sure how much philosophy there was to take away from it. I managed to go to hear Strawson and Ryle only once or twice each. I did philosophy of religion as a 'special subject', however, for my theology degree at Oxford. I went to all of Ian Ramsey's graduate classes in philosophy of religion. His approach to reconciling theology with logical empiricism was hopeless, but he was a hugely generous sponsor of stimulating and valuable discussion. And at Princeton Seminary I was fortunate to have a philosophy of religion seminar taught by John Hick, as I was at Oxford to attend Austin Farrer's lectures, for two terms, on philosophical topics in Thomas Aquinas' theology. I consider them two of the most outstanding philosophical theologians who have been at work during my lifetime, very different in their approaches; and it has been a privilege and an inspiration to know Hick over the years since then.

John Marsh, my tutor in philosophy of religion at Oxford, got me working on Anselm's so-called ontological argument for the existence of God. I noticed the modal form of the argument in Anselm's response to Gaunilo, and was intrigued by it, but at the time I couldn't find out enough about modal logic to do much with it. John Hick encouraged me to keep working on the argument, and pointed me to Charles Hartshorne's work on it, which contained the basics of the relevant modal logic; and I put in quite a bit of effort studying that during my years in Montauk.

IV

In 1965, roughly as I had planned, after three years in the pastorate, I became a student again, in the Ph.D. program in philosophy at Cornell University. The chair of the Cornell philosophy department at the time was Norman Malcolm, and his ordinary-language-based Wittgensteinianism was the dominant influence in it. In philosophical methodology, that influence did

not prevail over the more Carnapian formation I had received at Princeton. I did welcome the loosening of the grip of empiricism on analytical philosophy, but did not think it needed to take a Wittgensteinian form. What I appreciate most about my education at Cornell was the unremitting demand for clarity and rigor in thinking and writing. When I arrived at Cornell, with rather grandiose ideas about what I might accomplish and how quickly, I had hardly begun to realize how hard philosophy is. When I left three years later, I had a much better understanding of that most important philosophical lesson.

A main reason why I went to Cornell was that Nelson Pike was teaching philosophy of religion there. He was a great encouragement, both in the seriousness with which he took theological and metaphysical ideas, and in his insistence that they be treated with clarity and rigor. He was also a great adviser, supportive and accessible, wise about philosophical strategies, demanding good philosophizing but not agreement in views. I considered writing a dissertation about the relation between religion and ethics, which was really the subject that most interested me. A very good ethics course I took as an undergraduate at Princeton, from Douglas Arner, got me thinking about it, and I had thought a lot about it at Oxford. But I had written little or nothing about it, and had much less worked out about it than I had on the ontological argument. That led me to conclude that it would be wiser, with a view to finishing my degree in good time, to write on a modal form of the ontological argument. I did that, and I did indeed manage to leave Cornell with a practically finished dissertation after only three years there. Within a few years I had quarried the dissertation for one published article and a significant part of another. But while my dissertation was thoroughly competent and (I still believe) largely correct, I have never felt there was enough important news in it to warrant working the whole of it up for publication as a real book.

The most obviously important thing I got out of my work on the dissertation, besides the timely completion of a Ph.D., was a pretty good grounding in modal logic and metaphysical issues related to it. I was essentially self-taught in modal logic, as I had been at Princeton in Kant. I knew no one at Cornell who knew as much about modal logic as I did, except Arthur Fine, who had just arrived to teach philosophy of science; it was helpful to check my understanding of it with him. The closest I found to a usable textbook in modal logic was Arthur Prior's philosophically

admirable *Formal Logic*, with its difficult Polish notation; the textbook by Hughes and Cresswell was not quite out yet. But it was clear to me that the literature on the subject was growing rapidly and modal logic was opening up as a very exciting field. It would be 'where the action was' in the 1970s, and my dissertation work left me prepared to have a bit of the action.

Of possibly greater, though less obvious, significance for my philosophical biography was the largest positive conclusion to which I found myself tending in my reflections on the modal argument for theism. That argument never seemed to me likely to persuade anyone of the existence of God, because any doubts about its theistic conclusion so easily turn into doubts about its premises. But in reflecting more broadly on issues of necessary existence I found myself drawn to the view that in thinking about logic and mathematics we are tracing structures whose existence is as necessary as the truths of logic and mathematics. Thinking about what sort of being those structures could have, I was drawn further to the thought that they are structures of God's thinking. I drafted a chapter on the argument for God's existence that Leibniz had based on this thought. In the end I did not include it in the dissertation, perhaps for the good and sufficient reason that the dissertation was about a different argument; or perhaps I was not yet ready to go so far out on that metaphysical limb. A quarter of a century later I did include in my Leibniz book a chapter on the argument 'from the reality of eternal truths'; and the argument, and the sort of theistic Platonism it represents, figure prominently in work that engages me still.

V

In 1968, I left Cornell to take up my first full-time faculty position at the University of Michigan in Ann Arbor. My four years at Michigan were pivotal in my philosophical formation. I don't think that after only three years in graduate school I had fully become a professional philosopher. I think I had by the time Marilyn and I moved to UCLA in 1972. I'm not sure that six- and seven-year Ph.D. programs, now the de facto norm, are desirable; but we will have them as long as academic employers prefer fully formed professionals for entry-level jobs.

Michigan hired me primarily to teach the history of modern philosophy—specifically, the seventeenth and eighteenth centuries; and that

teaching proved to be, philosophically, the most formative part of my experience at Michigan. Relative to my own interests and sense of vocation, it was also one of the more contingent turning-points in my philosophical biography. I had taken a lot of courses in the history of philosophy at Cornell, and certainly considered it one of the things I was prepared to teach, and interested in teaching. A seminar on Locke, Berkeley, and Hume that Jonathan Bennett taught as a visitor at Cornell had reawakened my interest in Berkeley; and my serious interest in Leibniz began in a seminar on him that Norman Malcolm taught. Malcolm was at his best on Leibniz; he really wanted to understand the great philosopher, and the material did not engage the intolerant rigidity that too often emerged when Wittgenstein was in view. But I had done even more work on ancient philosophy, and thought I was as ready to teach ancient as early modern. And of course my number one specialization, my dissertation field, was philosophy of religion.

Michigan already had a philosopher of religion, one of the leaders of the field, George Mavrodes. They were willing for me to spend half my teaching time in philosophy of religion, or any other field of philosophy in which I might be interested and competent. But what they really wanted me to teach was the history of seventeenth- and eighteenth-century philosophy, and they were insistent that I should spend half my teaching time in that field. Finding the Michigan philosophy department very attractive, I took the job and committed myself to the teaching in early modern. I have never regretted it.

About half of the teaching I have done in my career as a whole, including a majority of my doctoral dissertation advising, has been in the history of modern philosophy. The field did not loom so large in my research plans at first. Almost all the writing that I did in it before the late 1980s began as lecture notes for teaching; but I eventually published a whole book on Leibniz.

It was quite specifically part of my job at Michigan to teach the one-semester survey course on early modern philosophy that was required of all undergraduate philosophy majors. I had very largely to invent the course for myself, as I had not taken any course that was based on the views I was coming to have of the structure of the history to be studied. Planning the course and preparing the lectures the first time I taught it was an enormous effort; I have never worked harder than I did that semester.

The work was also enormously rewarding, and a major part of my own education in philosophy. In the late 1960s, analytical philosophers who wanted to think about metaphysics were still struggling to figure out how to do it. I found that the great philosophers of the seventeenth and eighteenth centuries had engaged metaphysical questions quite directly, and had done so, in their best work, with the sort of clarity and rigor to which analytical philosophers aspired. They became my models for thinking about metaphysics and, in effect, my teachers in metaphysics.

I think that teaching the history of philosophy has also had more general effects on my conception of philosophy. The canonical figures in a survey of early modern philosophy were systematic philosophers. Their systematicity is one of the attractions that has kept them in our canon. We would like to be able to put the world together in our minds, and we are interested in ways of trying to do it. As Tyler Burge remarked to me years ago at UCLA, it is an attraction of teaching the history of philosophy that it offers the chance to expound and discuss a large philosophical system (or more than one of them) even if one has not yet worked out a system of one's own.

Systematicity is not just an aspiration. As one studies the systems of great philosophers in the receptive but critical frame of mind that is necessary for getting the most out of them, one experiments with them, pulling a string here to see what moves over there, so to speak. What happens to the system as a whole if this thesis is dropped, or that implausible or clearly outdated doctrine is revised in one or another way? One learns that some doctrines can survive credibly without the system, and the system without them, and that others are not so detachable. In the process one discovers not only the internal connectedness of the views of this or that philosopher, but the intrinsic systematicity of the subject-matter, the interrelatedness of the problems of philosophy.

I had been trained in an 'article culture' that thought of analytical philosophy as 'piecemeal philosophy', a social project like the natural sciences, in which we are not trying to build our own individual systems, but each trying to contribute a bit here and a bit there to the progress of a cooperative intellectual enterprise. That model has surely been salutary in important ways for the health of our discipline. And, clearly, since none of us can do everything at once, it is important to learn to discern topics and issues that can be at least provisionally excluded from any philosophical project one is working on. But philosophy also resists piecemeal treatment. Except

for the most straightforwardly empirical facts that figure in philosophical reasoning, philosophical theses tend to be fragments of actual or potential systems, and to look quite different in different systematic contexts. I think that is one of the reasons why the undergraduate courses that seemed to me most successful, in my teaching of philosophy, have generally been courses that focused on great books of at least moderately large scope.

I believe it is also salutary that serious work in the history of philosophy leads one to think about major philosophical issues, not just in one way, but from the diverse points of view that are represented in the history one studies. A good philosophical understanding of a philosopher's work is never uncritical. We need to explore objections to the work in order to put the philosophy through its paces and discern its implications and motivations. But the aim of the philosophical historian's critical examination is not to determine what is the true theory or the best point of view. It may be healthy to try to make such judgments for ourselves; but our endorsement of them, as distinct from the arguments we contrive for them, is not likely to be a particularly important part of our professional contribution to the discipline.

That is largely because the progress of the discipline is not to be found in such judgments. Few of the big questions of philosophy have been permanently settled. Few of the main theoretical positions have been conclusively determined to be right or wrong. Philosophy has been much more successful in exploring possible ways of thinking, giving us a clearer, deeper, and fuller understanding of them, than in generating agreement as to which of those ways of thinking accord best with reality. It is plausible to think that will continue to be the case, because it is plausible to suppose that the contents and relations of philosophical views and questions are more accessible to us intellectually than many of the facts that would make the views true or false as representations of reality.

This is not to say that we should not expect philosophy to help us deal with reality. Even if we do not have agreed answers to large issues of metaphysics and metaethics, a philosophical understanding of concepts and arguments related to those issues may help us think in clearer-headed and uncontroversially better ways about particular scientific and ethical questions. But I do not think that is the deepest reason for studying philosophy and its history. The realm that philosophy is likeliest to succeed in exploring, the realm of possible ways of thinking, is full of objects of great beauty. It is worth loving for its own sake.

It is hard to date my falling in love with philosophy. It probably began in my undergraduate years, as I found in the clarity and rigor of analytical philosophy's formulations and arguments the same sort of beauty I had learned in high school to see in mathematical proofs. That is of course one of the forms of the experience, and love, of beauty that are celebrated in the speech of Diotima in Plato's *Symposium*. In the theistic Platonist's view it is also a glimpse of the beauty of the divine mind. I began to study philosophy, no doubt, with the thought of using it to serve other interests of a religious sort. But I have come to think that the deepest religious significance of philosophy demands that it be loved and practiced for its own sake.

VI

I think of three further ways in which the four years in Ann Arbor set directions for my future philosophical work. Two of them I will discuss rather briefly; the third will open a longer discussion in the next section. (1) During my theological studies and my years of ministry in Montauk I had devoted considerable time to reading nineteenth- and twentieth-century religious thinkers, of a generally 'Continental' philosophical orientation; but I arrived at Michigan with no intention of working further on them professionally. Quite likely I never would have, had it not been for the influence of Jack Meiland, a senior colleague at Michigan. He persuaded me of the pedagogical value of teaching such material to undergraduates, and in my second term in Ann Arbor I gave un undergraduate seminar on four nineteenth- and twentieth-century Continental religious thinkers. I continued to teach this material throughout my career, and greatly enjoyed the way it engaged undergraduates' interests. At Michigan and UCLA I did not find much graduate student appetite for courses in this field; one of the things I enjoyed, much later, about my situation at Yale was the opportunity to teach seminars on Schleiermacher to groups composed of doctoral students in theology as well as undergraduates. This area has not been a main focus of my research, but over the years I have published about half a dozen essays on Schleiermacher, Kierkegaard, and Buber.

(2) One of the doctoral students I had the good fortune to advise at Michigan was William A. Polkowski. I learned a great deal from his thesis

research on 'The Possible Evidential Value of Religious Experience', and particularly from his use of Bayes' Theorem in the calculus of probabilities, which he had studied at Michigan with Arthur Burks. I became very interested in the relevance of Bayesian considerations to metaphysics and epistemology. Seeing the ineliminable place of 'prior' assignments of absolute and conditional probability in Bayesian reasoning helped to make clear to me that I should not any longer count myself as an empiricist about the justification of belief.

VII

During my first term in Ann Arbor I taught a topical survey of the philosophy of religion as an undergraduate lecture course. It was a pedagogical disaster, pitched way over the students' heads; but a lot of my later work grew out of it. In it I began to open up the topics in the relation between religion and ethics that I had prudently set aside at Cornell in order to write a dissertation I could finish quickly. One of these topics was the divine-command theory of the nature of moral obligation. My first published essay on that subject, 'A Modified Divine Command Theory of Ethical Wrongness', was written at Michigan in response to an invitation obtained for me by my senior colleague Bill Frankena to contribute a paper to an anthology on religion and ethics. In my own view none of my contributions to philosophy is more significant than the work I have done, beginning with that essay, towards the development of a viable theistic metaethics.

The development of my own position on the subject has not exactly followed a direct path. I published a few further essays on divine-command theories, casting them in different lights. But they did not add up, in my own opinion, to a complete metaethics, because they presented only theories of obligation, or of right and wrong, and I was not ready to offer a theory of the good.

I had some thought that a theistic metaethical theory of the good might be sought in reflection about God's goodness and love, a subject which interested me also in relation to work I was doing on the problem of evil. Accordingly, during my first year at UCLA, in 1972–3, I began writing about the nature and ethical significance of love, both divine and human;

and I conceived the project of writing a book on the subject. That was my project for 1974–5, the first year that Marilyn and I took leave from UCLA, with the aid of fellowships from the National Endowment for the Humanities. We spent the year in Oxford, and it was immensely fruitful for me. The reading I did there, and discussions I had, especially with Derek Parfit, laid major foundations for all the subsequent work I have done in ethical theory. I also drafted several chapters on love; but it was clear to me at the end of the year that they were not adding up to a book, and I was not ready to publish them. The only piece from the project that I published in the immediate aftermath of the leave was my article 'Motive Utilitarianism'. In relation to my larger project, it was originally conceived only as a prolegomenon, defending the independent significance of the ethics of attitudes, or more broadly of 'agent ethics', as distinct from the ethics of actions. I went on for years teaching classes and seminars on the ethics of love, but it took further catalysts to bring my ideas on the subject into a synthesis (a larger synthesis) that I found satisfying.

Two catalysts stand out in my memory. One was an invitation to give the Wilde Lectures on Natural Religion in Oxford. I accepted, with the plan of giving them on the relation between religion and ethics, committing myself to something more like a book on the subject. The other catalyst was supplied in a discussion I had, late at night at a conference during the 1980s, with Bill Alston and Al Plantinga, in which they pressed on me the question why I should not think that the goodness of things is to be understood in terms of resemblance to God. I don't remember with confidence how the discussion started, but I think it was connected with thinking Bill had been doing on the relation between Platonic metaethics and theistic metaethics. As I planned my Wilde Lectures I became more and more interested in a theistic Platonism in which God occupies something of the role that the form of the Good (or of Beauty) occupies in Plato's 'middle dialogues', and more and more convinced of the centrality of the idea of intrinsic excellence, both for ethical theory and for theology.

The Wilde Lectures that I gave in Oxford in the spring of 1989 started with those ideas, and developed a metaethical view that gives the idea of excellence priority in relation to the idea of obligation. I worked on the lectures for practically ten years more (alongside other projects) before I finally had a book on the subject. Our move to Yale in 1993 provided a helpful situation for this work, one in which I had occasion to teach more

in ethical theory than I had before. It was also helpful that Yale's lively interdisciplinary culture put me in conversation about ethics with students and colleagues in theology and religious studies, political theory, and law, as well as philosophy. When *Finite and Infinite Goods* came out in 1999, the metaethical theory presented in it also provided a context in which (as I had realized only in 1997) much of the work I had been doing on the ethics of love could form part of a coherent whole. Not that it completed the development of my views in agent ethics. I left the nature of virtue somewhat to the side as a topic in that book, but have focused on it more recently, writing *A Theory of Virtue*, published in 2006.

VIII

One main context for my thinking about the relation between religion and ethics has been the Society of Christian Philosophers, which a number of us formed in 1978 with a view to helping and encouraging each other to integrate our Christian faith and our philosophical vocation. It has certainly helped and encouraged me to do that. Personal integration is a difficult business in any case, and the integration of personal identity as a religious believer and as a philosopher is particularly delicate. Not that I have ever seen philosophy and religious belief as inherently opposed. On the contrary, in common with major traditions in the world's most developed religions, I believe that religious thought, and even spiritual meditation, can advantageously take a philosophical form. But even where faith and philosophy are married, each has its own integrity, and there will be tensions. It requires some courage for the believer to acquire the experience that teaches the limits of what philosophy can do either for or to religion. And it is a potentially crippling temptation for religious philosophers to adopt a primarily defensive and protective stance in relation to religious doctrines, where what is really needed is creative and imaginative thinking about religious questions.

I do not believe in drawing a sharp line between philosophy and theology. Especially in ethics I think one ought to bring one's whole self to one's thinking. What I have written in moral philosophy since the early 1980s has certainly been influenced by Christian beliefs and sources, and has sometimes touched quite explicitly on theological themes and issues. At the

same time I have usually written for a general philosophical audience. In that context I have not wished to presuppose commitment to Christianity, and I hope that Christian ideas may shed light on ethical views that will commend themselves also to people who are not Christians.

Philosophy of religion is among the areas that have benefited most from the tremendous development and broadening of analytical philosophy in the last half century. When I began to study the subject in the 1950s, I could easily carry in my hands the small pile of volumes containing practically all that had then been written in contemporary analytical philosophy about religious issues. On the whole it was not a very satisfying library. Today there is a large analytical literature in the philosophy of religion, of the sort that I wanted to read, and rarely found, in my student years. I am pleased to have been able to contribute something to that development.

IX

More broadly, I am proud of the contributions of my generation in analytical philosophy. To find a fit comparison for the flowering of rigorous philosophizing, in English, on an ever-widening range of topics and ideas, in a context of mutually illuminating discussion, in the last 100 and especially the last fifty years or so, one might have to go back to the thirteenth and fourteenth centuries. For me the most exciting philosophical development in my adult lifetime was the explosion of interest and activity in the 1970s connected with the so-called 'new theory of reference' (or *direct* reference) and possible world semantics for modal logic. Those developments began, of course, in logic and philosophy of language, but I was most interested in possibilities they opened up for metaphysics. The idea they generated that has stayed with me the longest (and eventually became central for my treatment of metaethics too) is that of ways of separating questions about the natures of things from questions of meaning.

I began working in this area during my last year or so at Michigan, where engagement with David Lewis's paper on 'Anselm and Actuality' led to my writing a first draft of my paper on 'Theories of Actuality'. After we moved to UCLA, the philosophical atmosphere there, and especially discussions with David Kaplan, were a great stimulus to further work on modal metaphysics and related issues about identity. I continued writing in

the area, and expected during our second leave from UCLA, in 1979 and 1980, to produce a book of which a major part would develop my views on these subjects.

Once again I found that I was not ready to write a projected book. That was largely due to the fact that the project was not simply to write a book about modal metaphysics, but to solve the theological problem of evil, using ideas about identity and its modality as a central part of the machinery for doing so. During the 1970s, I had published a couple of papers based on the Leibnizian thought that if evil had never existed, you and I would never have existed either, but, at most, as other individuals similar to us. I initially saw that idea as offering a promising framework for theodicy, but during 1979 and 1980 I was forced to conclude that it would not solve enough of the problem. The only thing I published in the project area as a result of that leave was my paper on 'Actualism and Thisness'.

I still think the Leibnizian idea about the connection between evils and our identity is relevant for thinking (and feeling) about the problem of evil. And I have recently published another, somewhat chastened and (I hope) better focused paper on that subject. But I doubt that I will publish a book on the problem of evil. Marilyn McCord Adams, my wife, has provided a much better framework for thinking about the problem. I believe her book, *Horrendous Evils and the Goodness of God*, is the deepest and most satisfying treatment of this inherently unsatisfying subject that we are likely to have any time soon.

About ten years ago, after a number of years of focusing on other areas, I began to work again on analytical metaphysics, this time with a primary focus on ontology, and no special connection to the problem of evil. When I wrote about actuality in the 1970s, I intended to go on to write also about existence, which I regarded, and still regard, as a distinct topic from that of actuality. I have been writing about existence, and related issues about substance; and I related them to ideas about God as 'being itself', in a series of four Gifford Lectures on 'God and Being' that I gave at the University of St Andrews in 1999.

The thought that connected the topics of God and being in my Gifford Lectures connects my current interests in ontology also with my theistic Platonist metaethics. Very much as I think of the goodness of other things as (very imperfect) imitation of God, theistic Platonists in the medieval and early modern periods tended to think that all the fundamental attributes

of finite things are imperfect imitations of attributes of God. That idea is the basis of the conception of God as *ens perfectissimum* or *ens realissimum* in the writings of Leibniz and Kant, for example. I have been thinking and writing about the prospects for that way of conceiving the relation between God and the structure of finite things. I hope to produce a book of metaphysics, but it remains to be seen just what form it may take.

X

I conclude this essay with the same summary of my philosophical convictions that concluded my autobiographical remarks at the conference at Yale. I believe that there is a metaphysically significant difference between appearance and reality; that there is a capital '*R*' Reality that grounds everything that appears; that it is mental; that it is good; and that doing philosophy can be a way of loving it.

2

Yet Another Anti-Molinist Argument*

DEAN ZIMMERMAN

I. Motivating Molinism

Introduction

'Molinism', in contemporary usage, is the name for a theory about the workings of divine providence. Its defenders include some of the most prominent contemporary Protestant and Catholic philosophical theologians.[1] Molinism is often said to be the only way to steer a middle course between two extremes: the radically opposed conceptions of foreknowledge, providence, and grace associated with Open Theism and Calvinism.

* I have benefited from the comments and criticisms of an embarrassingly large number of philosophers: at the 2004 Wheaton Philosophy Conference, where the argument was first presented; at the Yale conference honoring Robert Adams; in a philosophy of religion seminar at Rutgers University; and at a meeting of the Joseph Butler Society in Oriel College, Oxford. I was encouraged to discover that Robin Collins had come up with a similar argument, quite independently. I owe especial debts to Josh Armstrong, William Lane Craig, Keith DeRose, Tom Flint, Daniel Fogal, John Hawthorne, David Hunt, Sam Newlands, Calvin Normore, Alex Pruss, Mike Rea, and Jason Turner; but I know I am forgetting someone, and that I have not even done justice to all of the objections I *do* remember.

[1] Among philosophical theologians based in the philosophy departments of Anglophone universities, Molinism may well be the most popular of five or six competing theories. For some defenses of Molinism, see Alvin Plantinga, 'Replies to My Colleagues', in J. Tomberlin and P. van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel, 1985), 313–96; Jonathan Kvanvig, *The Possibility of an All-Knowing God* (New York: St Martin's Press, 1986); Richard Otte, 'A Defense of Middle Knowledge', *International Journal for the Philosophy of Religion*, 21 (1987), 161–9; Alfred J. Freddoso, Introduction, *Luis de Molina: On Divine Foreknowledge* (Part IV of the *Concordia*), trans. Freddoso (Ithaca, NY and London: Cornell University Press, 1988), 1–81; Edward Wierenga, *The Nature of God* (Ithaca, NY and London: Cornell University Press, 1989); and Thomas P. Flint, *Divine Providence: The Molinist Account* (Ithaca, NY and London: Cornell University Press, 1998).

Robert Adams, William Hasker, and others have formulated powerful arguments against Molinism.[2] I believe their work has uncovered a deep problem with Molinism: it posits 'brute' or 'ungrounded' facts concerning matters that require 'grounding' in more fundamental facts. The argument I develop against Molinism is in some respects less illuminating than theirs; it does not throw Molinism's deepest problems into relief. In another way, however, it is slightly more ambitious. Molinist feathers are often unruffled by complaints about 'ungrounded' facts and the apparent 'explanatory circularities' to which they lead. Groundedness and bruteness are metaphysically loaded notions; they—and the principles alleged, by anti-Molinists, to govern them—are complex and contested; Molinists have found ways to cast doubt upon their deployment in the arguments of Adams and company.[3] I try to show that Molinism has highly unintuitive consequences that are independent of grounding worries.

I begin with a rough sketch of Open Theism and Calvinism, highlighting the problematic aspects of each view, and the way in which Molinism is supposed to avoid them, serving as a mean between two theological extremes. The background is intended merely to explain why Molinism is important, and why so many contemporary philosophers and theologians have little alternative but to accept the doctrine. Readers familiar with Molinism and already convinced of its importance may wish to skip ahead to section II.

Alternatives to Molinism: Open Theism and Calvinism

Open Theists are libertarians; they think that we would not be free if our decisions were the inevitable outcome of the distant past or God's

[2] Cf. Robert M. Adams, 'Middle Knowledge and the Problem of Evil', *American Philosophical Quarterly*, 14 (1977), 109–17; and id., 'An Anti-Molinist Argument', in *Philosophical Perspectives*, v, ed. J. Tomberlin (Atascadero, Calif.: Ridgeview, 1991), 343–53; William Hasker, *God, Time, and Knowledge* (Ithaca, NY and London: Cornell University Press, 1989); id., 'A New Anti-Molinist Argument', *Religious Studies*, 35 (1999), 291–7; David P. Hunt, 'Middle Knowledge: The "Foreknowledge Defense"', *International Journal for Philosophy of Religion*, 28 (1990), 1–24; and Timothy O'Connor, 'The Impossibility of Middle Knowledge', *Philosophical Studies*, 66 (1992), 139–66.

[3] Flint responds to numerous versions of the 'grounding objection' in *Divine Providence*, chs. 5 and 6. Adams's 'An Anti-Molinist Argument' turns upon a transitive relation of 'explanatory priority'. Flint argues that it is not obvious that the same relation is being invoked each time Adams appeals to explanatory priority; and that, if it is the same relation, it is not obviously transitive. Cf. Thomas P. Flint, 'A New Anti-Anti-Molinist Argument', *Religious Studies*, 35 (1999), 299–305, and id., *Divine Providence*, ch. 7.

irresistible, prior decrees.[4] What exactly is meant by 'free' in this context is a nice question; but the libertarians who are involved in this debate generally assume there is an important variety of freedom that is incompatible with determinism, necessary for moral responsibility, and usually implicated in serious assertions that some event was 'up to me' or 'within my power'. Many Christians have suspected that a good deal of the evil God permits in our world (perhaps, indirectly, all of it) is due to the fact that there is some great value in creating genuinely free and responsible creatures—persons whose choices God cannot simply determine, without abrogating their freedom and making them no longer responsible for their actions. This much of the Open Theist agenda enjoys wide support. More radically, however, Open Theists think freedom requires that the future be 'genuinely open'—that there be no fact of the matter, ahead of time, about what I will freely choose. But, in that case, there is no fact for God to know, ahead of time.[5] The amount of providential control God exercises over creation is limited by the extent to which he[6] leaves the future open to the influence of our free decisions (and whatever other genuinely 'chancy' processes he might allow[7]).

The amount of 'openness' Open Theists need is a matter of some controversy among them. Of course, God knows precisely which alternatives

[4] For detailed defense of Open Theism on philosophical and theological grounds, see Hasker, *God, Time, and Knowledge*; Clark Pinnock, et al., *The Openness of God* (Downers Grove, Ill.: InterVarsity Press, 1994); David Basinger, *The Case for Freewill Theism: A Philosophical Assessment* (Downers Grove, Ill.: InterVarsity Press, 1996); John Sanders, *The God Who Risks: A Theology of Providence* (Downers Grove, Ill.: InterVarsity Press, 1998); and Gregory A. Boyd, *God of the Possible* (Grand Rapids, Mich.: Baker Books, 2000).

[5] Another theological position that belongs to the same family as Open Theism is the slightly more radical thesis that, although there *is* a fact of the matter about what I will do, God does not know it ahead of time. Richard Swinburne and Peter van Inwagen hold this view because they believe that God could not know what I will do unless it were inevitable; and that the sort of inevitability that would be required for God to know it is incompatible with freedom. See Richard Swinburne, *The Coherence of Theism*, rev. edn. (Oxford: Clarendon Press, 1993), 167–83, and Peter van Inwagen, 'What Does an Omniscient Being Know about the Future?', in Jonathan Kvanvig (ed.), *Oxford Studies in Philosophy of Religion*, i (Oxford: Oxford University Press, 2008).

[6] My use of the masculine pronoun when referring to the deity is a sign of conservatism in matters of English style, not theology. It strikes me as absurd to use the feminine pronoun when referring to the undoubtedly male Jesus Christ; but, beyond that, I see no compelling *theological* argument (on general Christian principles) for the inevitability or importance of using only masculine pronouns when referring to God. Attributing masculinity to God is metaphorical at best; and the Hebrew and Christian scriptures use both feminine and masculine metaphors to describe God.

[7] Van Inwagen believes God may have left a great deal up to chance besides our free choices. See his 'The Place of Chance in a World Sustained by God', in Thomas Morris (ed.), *Divine and Human Action* (Ithaca, NY and London: Cornell University Press, 1988), 211–35.

he has left genuinely open (perhaps some that seem to us to be live options are really not); and God knows the range of responses he could make in the future, as the story of his relationship with humanity unfolds. Furthermore, there is plenty of biblical and theological precedent for supposing that God sometimes *makes us do things* in ways that admittedly render us mere vehicles for God's actions, and therefore not personally responsible for what we do. So it is not as though the God of the Open Theists can never infallibly predict what someone will choose to do—just not what they will choose on those occasions when they are allowed to exercise genuine freedom. It need be no part of this picture of divine providence that God is ever *surprised* by the outcomes of the decisions he leaves up to us. But it does involve his taking *risks*: God may know 'the end from the beginning', because he can see that all the genuinely open alternatives can be made to converge, in one way or another, upon an outcome that God chooses. Still, according to Open Theists, between creation and eschaton, God allows many situations to develop without his having prior knowledge of exactly how they will turn out.

The Open Theists' picture of foreknowledge and providence includes two theses that conflict with Catholic teaching and most Protestant theological traditions. Open Theism may save the letter of the traditional doctrine of God's omniscience—God can know all truths, and yet not know what will happen, so long as there is now no fact of the matter about what will happen. Still, most Christians have affirmed something the Open Theist denies: that God has knowledge, at all times (or perhaps from a timeless perspective), of everything that will ever occur. Secondly, Open Theists embrace a 'risky' conception of the way God guides the course of history: God makes the decision to allow a certain course of events to unfold *before* he knows exactly what the outcome will be.

Far to the other side of the spectrum from the Open Theists and their view of providence there are Christians like John Calvin who think that I can be morally responsible for a voluntary decision, despite the fact that God caused me to make that choice. If determinism is true, God set up a chain of cause-and-effect starting as far back as the Big Bang, including a series of events that led inevitably to this decision. Or, even if he left the decision-making process 'indeterministic', from the point of view of natural laws; nevertheless, he may have determined its outcome, in advance, by divine decree. Of course, if all choices are caused in one of these ways, there

would be no reason to doubt that, from all eternity, God knew exactly what would happen in the course of human history, so long as he knew what he, himself, would choose to do; nor would there be any mystery about how God could insure that history take the course he desires. I shall call this kind of divine determinism about providence 'Calvinism'—though Calvin had distinctive things to say about many other matters, and I am glossing over subtle differences amongst Calvinists concerning the degree to which our choices are thought to be predetermined.

Calvinistic theology seems to be growing in popularity, at least among conservative Protestant intellectuals in North America.[8] But it is not for everyone. It will not appeal to Christians who hope to hew closely to orthodoxy within churches and theological traditions that come down on the side of Arminius rather than Calvin. And increased enthusiasm for Calvinism is not detectable within philosophy. It appears to me that most Christian philosophers—including many who, like Nicholas Wolterstorff and Alvin Plantinga, identify closely with Calvinist theological traditions—reject Calvin's teachings on grace and predestination.

Why does Calvinism have much less appeal for Christian philosophers than theologians? No doubt there are many factors at work. One that seems salient is the fact that most Christian philosophers receive their training and do their teaching surrounded by people who think the problem of evil decisively disproves the content of their faith; and we are routinely required to explain how we can maintain belief 'in the teeth of the evidence'. Libertarian theories of freedom provide a means for us philosophers to explain what the point of a great deal of evil *might* be, and in a way that at least makes some kind of sense to our largely skeptical colleagues and students. Even philosophers who reject libertarianism can see the internal logic of the explanation. Christian intellectuals based in less hostile territory no doubt encounter just as much evil, and probably spend as much time worrying about the problem of evil. But the mentors, peers, and students of theologians and church leaders do not take the problem of evil to be a knock-down argument for atheism—an argument so strong that only the

[8] A large proportion of American Evangelical churches can trace their roots to Wesley via Pentecostalism or the Holiness Movement—all staunchly Arminian—but anecdotal evidence suggests that many leaders within these churches are attempting to steer their flocks away from Wesley and towards Calvin. The battles between Calvinistic and Arminian Baptists go back to the earliest days of their movement; but, today, the Baptists' largest denominations and loudest voices side with Calvin. For a battlefield report, see Colin Hansen, 'Young, Restless, Reformed', *Christianity Today* (Sept. 2006), 32–8.

dim-witted or intellectually dishonest could doubt its soundness. And that is what many of us philosophers have been up against.

If this difference in our cultural milieus does partly explain Calvinism's unpopularity among philosophers and popularity among the Christian intellectual leadership outside philosophy, this need not be taken to show that philosophers are somehow better placed to know the truth. The God of Calvinism does not strike the people in my environment as a being who loves all his creatures and is truly worthy of worship. Calvinists may say (in fact, have said, in the blogosphere!) that the fact that I heartily endorse this reaction (and, for the record, I do endorse it) merely shows the extent to which my thinking conforms to the standards of 'the world', as opposed to those of true Christianity. The idea is not wholly implausible: philosophers with a Calvinist heritage who embrace libertarianism have simply been driven into apostasy by the greater pressure to explain themselves; and those of us philosophers who identify with traditionally Arminian theological traditions would see the superiority of Calvinism, as many of our best theologians have done, were we not so sensitive to the ambient skepticism.[9]

Whether for good reasons or bad, most Christian philosophers find themselves in search of a middle way between these extremes. They want a theory of providence that allows for libertarianism about free will (and libertarianism of a sort that helps to explain the existence of moral evil); but a theory that also affirms complete foreknowledge and rejects the Open Theists' 'risky' view of providence. Molinism's contemporary defenders present their view as an essential part of a doctrine of divine providence that can meet these desiderata; and they often allege, quite plausibly, that it is the only theory that can do the trick.[10]

II. The Molinist's Theory of Foreknowledge

Foreknowledge and 'Deep Explanations' for Actions

There are very general arguments for the incompatibility of our freedom with divine foreknowledge (or even with complete knowledge, from a

[9] Keith DeRose quoted me on this issue in a weblog, and at least one Calvinist scholar gave this spin to my explanation.

[10] e.g., Flint, *Divine Providence,* ch. 3.

timeless perspective, of what is future relative to us).[11] But let us assume that they fail—that, so long as God leaves our choosing undetermined, and gives us whatever else a libertarian might think we need in order to have the freedom to choose from among a range of alternatives, then God's merely knowing about it ahead of time is no threat to freedom. (I find the arguments *against* this assumption rather impressive; but they will drive libertarians directly into the arms of the Open Theists; and, here, I am exploring the viability of 'middle ways'.)

For the purpose of comparing Molinism and its rivals, I shall generally assume that God can properly be said to exist, act, and know things contemporaneously with events in our universe (although I shall make occasional remarks about the case of an omniscient but timelessly eternal deity). I shall also assume that God existed prior to his creation of anything at all. The puzzles for God's freely choosing to create a world, while knowing everything about the history of that world, would arise even had God always coexisted with created things. But I will ignore the complexities this possibility would introduce.

Could God have chosen to create a universe of a certain type, for good reasons, while utilizing every bit of his foreknowledge (or timeless knowledge) in making this choice? Numerous puzzles have been raised for the combination of foreknowledge (or timeless omniscience) with rational choice. There is something strange about the idea of a person's choosing to make something happen when he already knows that it is going to happen; or the idea of his deliberating over something when he knows he is going to do it.

The difficulty of imagining *ourselves* in such situations should probably not be taken to indicate anything deeply problematic about combining divine foreknowledge with rational, free, divine choices. Even remaining on a crudely anthropomorphic level, we can make some sense of the combination. A God with foreknowledge is rather like a time traveler who circles back and meets herself; both have special knowledge about what they will do before they do it. The time traveler's younger self saw

[11] For a classic statement of such an argument, see Nelson Pike, 'Divine Omniscience and Voluntary Action', *Philosophical Review*, 74 (1965), 27–46. For discussion of a modified version targeting timeless omniscience, see Plantinga, 'On Ockham's Way Out', *Faith and Philosophy*, 3 (1986), 235–69, repr. in John Martin Fischer (ed.), *God, Foreknowledge, and Freedom* (Stanford, Calif.: Stanford University Press), 178–215 (citations refer to Fischer, 183–4).

her time-traveling older self doing certain things; and now, after growing older and going back in time, she remembers seeing herself do what she is about to do. One can tell stories in which it seems the time traveler could choose to do things for reasons that include her memory that she will do these things. For example, she might worry that, were she to choose to do something other than what she remembers, she would make it the case that contradictions are true, and then terrible things would happen. (Like the characters in the movie *Dogma*, she might worry that everything would cease to exist if she makes a contradiction true.) Could a person rationally believe such a thing? (With that question, the characters in *Dogma* are of little help.) If so, she would be rational in choosing to do what she remembers doing precisely because she remembers doing it—so a rational choice *could* be made on the basis of a reason that crucially includes knowledge of what choice will be made. Our time traveler might not need to believe anything quite so bizarre in order to choose on such a basis. Suppose she is simply a very passive person, someone who never wants to rock the boat; the fact that she knows that she did something at such-and-such time and place could be seen by her as a good reason to do it; perhaps in some cases the only reason.

Would the time traveler's knowledge be an obstacle to *deliberation* about the foreknown act or choice? The time traveler can certainly rehearse various reasons for and against doing something, including the fact that she remembers doing it. Would such inner rehearsal count as deliberation? Perhaps it would. Suppose she says: 'I considered whether or not to jump into the river to rescue the drowning man; and although I knew that I would do it (I distinctly remember, as a young girl, seeing my time-traveling older self diving into the river), and although I could have done it merely to "go along with the flow of history", in fact I did it out of compassion for the victim; one often has several beliefs that could serve as good reasons to do something, but not all of them need be the actual reason for which one acts.' I am not at all sure that I see anything deeply wrong with that little monologue; and it sounds rather like deliberation while having full foreknowledge of the decision to be made.

I do not, then, see an easy way to prove the impossibility of someone having complete foreknowledge, including knowledge of her own decisions, while nevertheless acting for reasons—reasons that may or may not include the foreknowledge she possesses about the act itself. Still, there is something

funny about all these cases. The time traveler who does what she does because she knows that is what she will do lacks a really satisfying explanation for her action. Worries that contradictions would be true, or the desire to 'go with the flow', may make the choice psychologically understandable. But ask her why the world contains that action rather than some other and she will draw a blank. Unless there is some sufficient causal explanation for the entire 'loop' including the action, her memory of it, and the decision to act, there is no further explanation to be given. Although it is hard to say anything uncontentious about the nature of explanation, the following principles sound pretty good to me: There can be a plausible *psychological explanation* of why a person chose to do such-and-such, even if the explanation appeals to the person having reasons that include knowledge that he had only because he *would choose* such-and-such; but, in these circumstances, there will be no truly *deep explanation* why the world contains both the knowledge and the choice, unless there is some independent explanation for both.

One need not accept the Principle of Sufficient Reason to think that there is something wrong with supposing that God takes major decisions without 'deep explanations'. Perhaps it is impious to think that God's reason for creating a red planet rather than a blue one was simply that he took a fancy to red planets; but far worse to say that he created a red planet rather than a blue one merely because he knew that is what he would do—for then he acts in ways not even he can explain. With respect to the important details of the creative act (or acts) by which God brought the universe into existence and holds it together, we should expect there to be deep explanations—explanations that do not, therefore, advert to foreknowledge of those very details.

'Stages' in God's Foreknowledge

The need for deep explanations of (at least some aspects of) God's creative choice leads the believer in complete foreknowledge (or timeless omniscience) to posit an ordering of the knowledge God has into various 'stages' or 'levels'. Some facts can serve among the reasons for God's making a world containing such-and-such, while others cannot. Relative to the decision to include such-and-such in the world, the facts that *can* play a role in explaining the decision come 'earlier than' those that *can't*—though not, of course, in any temporal sense. Christians have typically believed

that God did not *have to* create; in which case, unless he lacks a deep explanation why he created anything, there must be a subset of the things that God knows that informed this decision; and it must not include his knowledge that there will be anything at all, other than himself.

It is hard to understand how a being with complete foreknowledge could 'bracket' some of it, acting only on the basis of part of what he knows—hard, but not hopelessly so. One simple-minded analogy appeals to what happens to *us* when things that we know slip our minds. If what is in fact knowledge that I will do *A* can be forgotten or ignored or bracketed somehow, then it becomes possible once again for me to choose between doing *A* and not doing *A* for reasons that are independent of my knowledge that I will do *A*. Imagine that I have been a 'passive' time traveler for many years, doing what I do simply because I remember doing it. Suddenly, I become tired of my passivity. I seek, instead, to 'live in the moment', ignoring what I know about my future while I am making decisions. If I succeed, my subsequent actions will be taken for reasons I have that are independent of my foreknowledge. Believers in complete foreknowledge (or complete timeless knowledge) must suppose that, in a roughly (no doubt *very* roughly) analogous way, God can ignore or somehow 'bracket' parts of what he knows, rendering them irrelevant to his decision to include this or that in his overall plan for the world. God's beliefs about what he will do, although they do not temporally succeed his choices about what to do, nevertheless 'come later in the order of explanation'. That God would freely choose to create Adam and Eve has always been known by God, but he has always known it because he has always 'already' chosen to create them; the choosing precedes, 'logically', the knowing. Anything that God knows, if it could serve among his reasons for so choosing, must come at a stage in God's knowledge that is prior to the knowledge that he would so choose.

Calvinists should readily agree that there are stages in God's knowledge. They merely need two such stages: (i) God's knowledge prior to his choice of a complete world, which consisted of his knowledge purged of the truths that depend upon his choice of a world—so, presumably, little more than necessary truths. And (ii) God's knowledge of everything whatsoever. The second stage follows hard upon God's choice of a complete history for the world, including every 'free' choice ever made by anyone. But libertarians who believe in complete foreknowledge have to say something much more complicated than Calvinists about the stages in God's knowledge.

I will use the label 'simple foreknowledge' for the following combination of views: God has complete foreknowledge, libertarianism is true, and Molinism is false.[12] Those who hold this combination of views must posit many stages in God's complete foreknowledge. (Libertarians who reject Molinism and accept divine timelessness will end up with a similar view: God's timeless knowledge must be divided into many stages.) Could God's decision to put a creature in certain circumstances be informed by his foreknowledge that the creature will in fact be in those circumstances and will choose one alternative rather than the other? It would seem not; for, if God's explanation for the decision included this fact, he would be unable to explain why the whole explanatory 'loop' exists: the creature's being in those circumstances, God's knowing that this would be the case, and his putting the creature in those circumstances based upon this knowledge. And so, according to the simple foreknowledge picture of the workings of providence, knowledge of what a creature will in fact freely do is not available at the stage prior to God's decision to create it and allow it to face this choice.[13] What distinguishes the Molinist from the believer in simple foreknowledge is the Molinist's willingness to say that there are truths of the form 'If creature *x* were in conditions *C*, *x* would freely do *A*'—conditionals that are not merely true because *x* will in fact be in *C* and will in fact freely do *A*. Rejecting Molinism requires that, if there are any true conditionals of that form, they are true because *x* will be in *C* and will then do *A*. True conditionals of the latter sort will generate explanatory 'loops', if they appear as crucial parts of God's reason to create *x* in *C*—and this would leave God without a deep explanation for his choice. Assuming

[12] David Hunt uses 'simple foreknowledge' for the conjunction of complete foreknowledge, libertarianism, and the thesis that one need not be a Molinist in order to believe the first two doctrines. But one needs a label for the stronger view, and 'simple foreknowledge' has been used for this as well—e.g., in the introduction to James K. Beilby and Paul R. Eddy (eds.), *Divine Foreknowledge: Four Views* (Downers Grove, Ill.: InterVarsity Press, 2001), 10. For a defense of simple foreknowledge, see Hunt, 'Divine Providence and Simple Foreknowledge', *Faith and Philosophy*, 10 (1993), 389–414; id., 'A Reply to My Critics', *Faith and Philosophy*, 10 (1993), 428–38; and id., 'The Simple-Foreknowledge View', in Beilby and Eddy, *Divine Foreknowledge*, 65–103. See also Bruce Reichenbach, 'God Limits His Power', in David Basinger and Randall Basinger (eds.), *Predestination and Free Will* (Downers Grove, Ill.: InterVarsity Press, 1986), 101–24.

[13] Although David Hunt argues that a God with simple foreknowledge would have more providential control than one without, I do not think he would disagree with this claim. See Hunt, 'Divine Providence and Simple Foreknowledge', and 'A Reply to My Critics'. My discussion of foreknowledge and 'stages' has been much improved by Hunt's insightful criticisms of earlier versions—though I fear he could still find things wrong with what I now say.

that God has deep explanations for creating each free creature and putting it in circumstances in which it exercises its freedom, and that the existence of later creatures and their circumstances often depend upon the outcomes of earlier free choices, the believer in simple foreknowledge must posit numerous stages in God's knowledge.

I pointed out two aspects of Open Theism that would strike many Christians as especially troubling: its denial of complete foreknowledge, and its 'risky' view of providence. The simple foreknowledge account just sketched (and its obvious analogue in the timeless case) can be faulted on the latter score. If stages prior to God's decision to create Adam and Eve are 'purged' of information that depends upon that decision, including facts about what they will do when tempted, then God takes a risk in creating them—he risks their succumbing to temptation, when (we may suppose) he hopes that they will not. And this is where the Molinist comes in, providing an alternative to both Open Theism and the simple foreknowledge picture of providence. Molinism posits a kind of information that satisfies two requirements: (i) it is available to God at stages prior to his deciding to create free agents, and (ii) it enables him to avoid *all* risks. Somehow, says the Molinist, God must know something about Adam and Eve that does not depend upon their existing and being tempted, but that nevertheless allows him to infer that, were they to be created and tempted in a certain way, then they would sin (or refrain from sin, as the case may be). The Molinist's solution is a simple one. There just *are* conditional facts of this sort, known by God, and true independently of the existence of Adam and Eve: If the pair were created and faced with such-and-such decisions, then they would freely choose to do so-and-so. With enough conditional facts of this sort available prior to any creative decisions, God need take no risks. The Molinist can claim other advantages, as well. When defenders of simple foreknowledge are asked to explain *how* God knows what will happen ahead of time, they are usually forced to say that it is just part of his nature to know everything. The Molinist, however, has a mechanism: God simply uses modus ponens. He considers the conditionals describing what creatures would freely do in various circumstances, decides what antecedents to make true, and infers consequents that add up to a complete description of all of history.

The Molinists' 'conditionals of freedom' (CFs) allow them to agree with the Calvinists about the number of stages in God's complete foreknowledge

(or, for Molinists who locate God outside time, stages in God's timeless knowledge): there are but two. The first stage consists of every fact that is independent of God's creative choices. These facts fall into two classes: (a) necessary truths and (b) CFs.[14] The information in (b) is exceedingly rich, according to the Molinists, allowing God to know exactly what choices would be made by every group of free creatures he could create, in every type of situation in which he could place them. (I shall go along with the common assumption that the same stock of possible individuals is available in every world. I favor a different view, but it would make no difference, ultimately, to the case against Molinism.[15]) Molina believed that CFs of *divine* freedom—i.e., conditionals specifying what God would freely do, given this or that set of CFs about possible creatures—are not known prior to God's decision to create, but are rather *chosen* by God as part of his one creative act. And most Molinists follow him in this.[16] With full knowledge of the true creaturely CFs, God simultaneously decides what

[14] Molinism acquired its other name ('the doctrine of middle knowledge') from Molina's contention that knowledge of (a) is, in the explanatory order of things, prior to knowledge of (b); and both (a) and (b) are explanatorily prior to God's complete foreknowledge, leaving (b) in the 'middle'. My more coarse-grained division ignores one of the distinctions in Molina's three-stage picture; but one can see how natural it is for the Molinist to regard CFs as being sandwiched between knowledge of necessary truths and the complete foreknowledge acquired at what I am calling the 'second stage'.

[15] If, as I suspect, 'singular truths' about individuals, including modal truths about them, depend for their existence upon the existence of the individuals that are their subject-matter; then we should adopt a modal logic like A. N. Prior's system Q; cf. Prior, *Time and Modality* (Oxford: Clarendon Press, 1957), ch. 5. In that case, what God knows, at stages before deciding to create anything, are purely general facts about what is possible for contingent individuals. And the Molinist should suppose that, for God to exercise risk-free providential control, he must know lots of CFs about the choices different person-types would freely take in various circumstances—with 'person-type' understood as a qualitatively specifiable role. In some of his earliest work on the problem of evil, Plantinga develops a Molinist theory of CFs involving 'possible persons' of this sort; see Alvin Plantinga, *God and Other Minds* (Ithaca, NY and London: Cornell University Press, 1967), 140–9.

[16] One might imagine that God decided what his own CFs would be prior to his knowing the CFs about creatures; in which case, the stages in God's knowledge would have to be ordered somewhat differently. (a), all by itself, would constitute the first stage; and, after a decision about which divine CFs should be true, the second stage would consist of (a) plus all CFs, both divine and creaturely; and complete foreknowledge would be inferable from this combination. (For discussion of this alternative, see Flint, *Divine Providence*, 55–65.) It is not clear that a Molinist picture of this sort would fully eliminate 'risk-taking', since God's decision about the divine CFs is made without taking into account the facts about creaturely CFs; and when God does take them into account, his choice of a world follows automatically. 'Before' knowing the creaturely CFs, or how the world would actually turn out, God made a decision; immediately, he knows the whole history of the world. In order to see whether this should be acceptable to someone of Molinist sympathies, one would have to undertake a close examination of the theological reasons to reject 'risky' views of providence.

Could a Molinist suppose that CFs of divine freedom are not chosen at all, but simply known by God along with CFs about creatures? I think not; for then all God's foreknowledge would collapse into a single stage, ruling out deep explanations for God's creative decisions.

he *would* do under the hypothesis that other CFs have been true, and also what he *will* do given this actual batch of CFs. God thereby decides what the world will be like in its entirety, start to finish, despite the presence of pockets of libertarian freedom. The Molinist need posit no more stages in God's foreknowledge (or timeless knowledge) than the Calvinist. There are necessary truths and CFs, constituting the first stage, prior to any creative decisions; then, after one gigantic creative choice on God's part—a choice that is enough, given the true CFs, to settle the whole history of the world, start to finish—there is God's complete foreknowledge.

As a good libertarian, the Molinist must say that the CFs are contingent. Were they not, then what I do in any given circumstances would be settled, ahead of time, as a matter of iron-clad necessity. Furthermore, as a good libertarian, the Molinist agrees that God cannot just make free creatures freely do whatever he wants. But if God could choose which CFs were true, he could do exactly that; so, creaturely CFs must be contingent truths over which God has no control. According to Molinism, then, it is as though God 'wakes up' to find certain contingent things true—there is an independent source of contingent fact at work 'before' God has a chance to do anything about it. Although Molinists may reject such talk as tendentiously impious, there is an important (and potentially troubling) truth behind it. The Molinist conditionals really are supposed to be contingent truths *discovered* by God, not determined by him; and discovered 'before' he creates—at least, 'before' in the order of explanatory priority. Thus, according to Molinism, if God wants to create free creatures, he does face certain limitations—despite the fact that he never actually 'takes risks'. God might turn out to be incredibly *unlucky* in the CFs with which he is forced to make do; although he does not *take* risks, he is nevertheless *subject to* risk.[17] This fact is important to latter-day Molinists, like Alvin Plantinga; it enables them to deploy the traditional 'Free Will Defense' against the problem of evil—and, in fact, to deploy it in a way that will ultimately prove important to my anti-Molinist argument.[18]

There is a striking contrast between the Molinist's use of the free will defense and that of the Open Theist or defender of simple foreknowledge.

[17] I thank Keith DeRose for this nice turn of phrase.

[18] For a statement of the Free Will Defense, under Molinist assumptions, see Alvin Plantinga, *God, Freedom, and Evil* (Grand Rapids, Mich.: Eerdmans, 1974), 7–64.

The Open Theist says God *literally* had to wait to see what I would freely do. He simply did not know, and could not know, what I would freely choose before he gave me the opportunity. So, how can he be blamed for allowing wrong choices, freely taken? The advocate of simple foreknowledge has the following to say about the origin of moral evil: God could not insure that I always (freely) do what is right, because he had to decide to create me and to put me in circumstances of free choice on the basis of only a part of his foreknowledge—a part that did not include knowledge of my actual choice. On the simple foreknowledge view, God does not have to 'wait to see what I will do' before he knows how things turn out, at least not *literally*; but, metaphorically, that is exactly what this sort of libertarian thinks God must do. Both Open Theist and simple-foreknowledge advocate say that God's decision to create free creatures was made under the risk of moral evil; but he had to make the decision despite the risk, if he wanted a world with free creatures and all the virtues that only free creatures can display. Obviously, some of us have badly abused our freedom; but, on either of these views, God had to give us opportunity to sin *before* (either literally or metaphorically) he knew that evil would result.

The Molinist, by contrast, denies that God ran any risk when he decided to create free agents. Nevertheless, since the CFs are contingent, and not under God's control, it is possible for them to prevent him from creating worlds he would very much like to have been able to create. God is dealt a certain set of CFs, says the Molinist; and he might find himself having to make the best of a very bad hand—so bad, that he simply could not create groups of free creatures facing significant moral dilemmas and always freely choosing well. (When the CFs about a certain possible creature turn out in this way, the creature has caught a bad case of what Plantinga calls 'transworld depravity', a syndrome to be described in more detail below.) Why does the Molinist think that every group of possible free creatures could have turned out to be 'transworld depraved'? It is assumed, at least by contemporary Molinists, that the way to generate the sets of CFs representing 'hands' God is 'dealt' in some possible world or other is by running through every consistent combination of CFs. This assumption, discussed in more detail below, will be essential to my argument against Molinism.

Another thing to notice at this point is that CFs are supposed to allow God to *avoid risk* and *maximize control* over creatures that nevertheless

remain genuinely free. If God knows what I would do when confronted with a certain sort of choice in a wide variety of circumstances, he can select the circumstances in which I would make the choice he most wants me to make, and avoid the ones in which I would make the choices he dislikes. The Molinistic theory of providence gives God much more control over me, and over the course of history as a whole, than the other two libertarian accounts of providence just described. This might seem to make the Molinist's God just as manipulative and coercive as the Calvinist's. But the Molinist will point out that God cannot just make us do whatever he likes; there is much about our free actions over which he has no control, due to his failure to be able to choose which CFs are true. Furthermore, the Molinist can plausibly maintain that, when God causes me to be in circumstances in which he knows I will freely do such-and-such, my going on to do such-and-such is not caused by God's putting me in those circumstances—at least, not in the more robust senses of 'causing' that are likely to threaten freedom. Granted, if one accepts a counterfactual theory of causation, and the CFs are counterfactuals, then this conclusion will be hard to avoid; but, otherwise, the Molinist ought to be able to say that God brings about a *necessary condition* of my choosing in the way I do, and it is only in that benign and uncontroversial sense that God can be said to cause my choice.

This description of Molinism and the motivations of its contemporary defenders should serve as a sufficiently detailed backdrop for the anti-Molinist argument to come.

III. The Conditionals of Freedom

Are CFs Counterfactuals, Subjunctives, or Something Else?

What kinds of conditionals are CFs? What conditionals will do the job for which Molinists need them? The examples I have used have been subjunctives, like 'If Eve were tempted, she would sin'; but that choice is not completely uncontroversial.

Plantinga called them 'counterfactuals of freedom' and the name has stuck. The name 'counterfactual' suggests that such conditionals must have antecedents that are 'contrary-to-fact'. But conditionals with *true*

antecedents must be among the CFs available when God decided whom to create and in what circumstances. Furthermore, it is tempting to say that, at that stage, it was 'not yet settled' which CFs would have true antecedents; and so 'not yet settled' which ones would be contrary-to-fact. In that case, none of the CFs known by God at the first stage would be a counterfactual—if 'counterfactual' really does mean 'contrary-to-fact'. Consider the conditional: 'Had Eve been tempted by a toad, she would not have sinned.' If its truth directly implies that Eve is never tempted by a toad, then its truth is presumably dependent upon God's not putting her in such circumstances; and in that case, it would not be available to God, prior to his decision to tempt her with a snake rather than a toad—at least, it is not something God knows at that stage, if there is to be any deep explanation of God's choice.

But perhaps the proposition expressed by this conditional sentence does not imply that Eve never be tempted by a toad; perhaps there is a more-or-less grammatical notion of 'counterfactual' that does not require that a true counterfactual have a false antecedent. As David Lewis has pointed out,[19] there are situations in which a conditional like, 'If Jones had been at the party, it would have raged until dawn', can be used to say something true, even though Jones *was* at the party. Usually when a person asserts something using this form of words, she expects the antecedent to be false; but perhaps such a statement can be true even when the expectation is not met. (Imagine the following response to someone who asserts the above counterfactual: 'What you said is true, but not for the reason you think; you see, unbeknownst to you, Jones arrived shortly after you left, and the party didn't fizzle out, like it seemed to be doing.') If we use 'counterfactual' to describe the grammatical and other linguistic features that distinguish these conditionals from other varieties (and not simply to mean or even to imply 'contrary-to-fact'), then Lewis's examples suggest that the CFs with true antecedents could be truly, albeit misleadingly, expressed as counterfactual conditionals.

But controversy over this question need not detain us. There are conditionals that will play the role Molinists assign to CFs, and that clearly need not have false antecedents to be true—namely, subjunctive conditionals. Suppose that, at a stage prior to God's decision to create Adam

[19] See David Lewis, *Counterfactuals* (Cambridge, Mass.: Harvard University Press, 1973), 26–8.

and Eve, the following subjunctive conditionals were true: (1) If Eve were tempted by a snake in such-and-such circumstances (ones that eventually came about), then she would sin; and (2) If Eve were tempted by a toad in such-and-such circumstances, then she would not sin. The Molinist who uses subjunctive conditionals for CFs can suppose that both were true, and available to serve among God's reasons for creating anything at all, let alone Eve and a snake. The second turned out to have a contrary-to-fact antecedent, and the former did not; so, if (contra Lewis) counterfactuals must be contrary-to-fact to be true, the Molinist can appeal to subjunctive conditionals instead of counterfactuals.

But must CFs be either counterfactuals or subjunctive conditionals? In English, at any rate, the only alternative is indicative conditionals, such as: If Eve is tempted by a snake, then she sins; and if Eve is tempted by a toad, then she does not sin. Could a Molinist plausibly claim that CFs are not subjunctives or counterfactuals, but indicative conditionals, instead? The only contemporary Molinist I know of who explicitly claims that CFs can be indicative conditionals is Richard Gaskin; but he thinks the indicatives in question have the same truth-conditions as closely related subjunctive conditionals, and he generally uses subjunctives as his paradigm cases.[20] At least one *opponent* of Molinism thinks the kinds of 'conditionals of deliberation' available to the Molinist's God should be construed as indicative rather than subjunctive.[21] Molinists have been happy with counterfactual or subjunctive CFs, and I will follow their lead. But, in an appendix, I argue that using indicatives as CFs would not help the Molinist to escape my argument.

I shall assume, then, that CFs are subjunctive or counterfactual conditionals—albeit ones that are rather unlike those we use to describe everyday events. Consider an ordinary subjunctive conditional: If I were to strike the match, it would light. This sort of claim will be true in some circumstances, false in others. The standard story about the truth conditions of such conditionals, due to Robert Stalnaker and David Lewis,

[20] See R. Gaskin, 'Conditionals of Freedom and Middle Knowledge', *Philosophical Quarterly*, 43 (1993), 414–16.

[21] Keith DeRose gives an account of counterfactuals and subjunctive conditionals that makes them more easily true than on many interpretations. If he is right, then relying upon them would not provide God with the kind of risk-free providential control the Molinist desires; so the Molinist should look elsewhere for conditionals to serve as CFs. See DeRose, 'The Conditionals of Deliberation', *Mind*, (forthcoming).

goes more or less as follows:[22] Take the actual world up to the time of the potential striking; change it just enough, if change is needed, to include the striking (and of course, if the conditional is contrary-to-fact, some change *will* be needed); and then 'see' whether the match lights in the 'nearest' world that results from following this recipe. How exactly to determine 'nearness' (which dimensions of similarity to weigh more heavily than others) is a vexed issue, as is the question whether one should assume that there always *is* a 'nearest world'. But two things seem clear enough: similarity of the laws of nature must play a particularly important role in determining nearness;[23] and differences that are later than the effect should almost always be ignored. A further common assumption, which the Molinist will question, is that the 'hypothetical' facts about a world, such as facts about which subjunctives are true there, must supervene upon the 'categorical' facts about the space of possible worlds. Suppose that, in the actual world, an opportunity arises for striking a match, and it is not taken. To figure out what would have happened, had the match been struck, one looks to possible worlds that have pasts very much like our world, but that are just different enough to include the striking. Take two such worlds, W1 and W2; in W1, the match lights, but in W2, it does not. It would be 'cheating', on the usual interpretation of the Stalnaker–Lewis semantics, to say that W1 is closer to actuality just in virtue of the fact that, in the actual world and in W1, it is true that, if the match were struck, it would light; but in W2, this conditional is obviously false. To appeal to this subjunctive similarity between W1 and the actual world would be to render the truth of this particular subjunctive 'brute'—it leaves a hypothetical fact not 'grounded' in the categorical. I will not attempt to say anything precise about the 'categorical'/'hypothetical' distinction, but simply help myself to the notion of 'sameness of categorical history'.

The standard way to apply the Stalnaker–Lewis approach to the case of the match would lead one to say things like: If the match is in fact wet, then, clearly, it would not light if struck, because the nearest world in which it is struck is one in which it is still wet. If oxygen has in fact

[22] For details, see Lewis, *Counterfactuals*; and Robert Stalnaker, 'A Theory of Conditionals', in Nicholas Rescher (ed.), 'Studies in Logical Theory', *American Philosophical Quarterly Monograph*, 2 (Oxford: Blackwell, 1968), 98–112.

[23] Exact sameness of laws is too much to require. In deterministic contexts, the worlds that seem most relevant to determining what would have happened are ones in which little miracles occurred in the recent past.

been evacuated from the room, again, the answer to 'Would it light, if it were struck?' is, no, since worlds with oxygen in the room are quite unlike the actual world. If, on the other hand, all conditions are right for lighting the match, so that the slightest scrape, together with laws like ours, imply combustion, then, yes, it would light if struck; in the nearest worlds with a match strike, its occurrence, plus the very similar laws and the relevantly similar conditions, together require that the world contain a lit match. Another possibility, however, is that the match does not do the same thing in all of the nearest worlds satisfying the hypothesis that I strike the match; that, in some of the nearest worlds, the match lights, while in others it does not. One might think this could only happen if determinism is false; but that is not so. Even if determinism happened to be true, our conditional claims would often turn out to be false—or at least not true—because of ties for the title 'nearest world' generated by the vagueness of ordinary language. Suppose the match head has very little inflammable material left on it, or that it is slightly damp on one side. There are some very specific ways of striking it that, together with certain actual, deterministic laws of nature, require its ignition, and others that require its failing to ignite. But the hypothesis that I strike the match is, inevitably, a rather vague one. We lack words for all the hyper-precise ways to strike a match, and some of the differences among these ways would matter, in this case. Suppose 'strike' is indeterminate between ways of striking that definitely would light the match and ways that definitely would not; and that nothing else about me or my situation decisively favors one of the successful strikings or one of the duds. In that case, if the actual world does not include the match's being struck on this occasion, the right answer to the question, 'Would the match light if I were to strike it?' would seem, again, to be *no*—or, at the very least, it should *not* be a definite *yes*.[24] In some of the nearest worlds the match lights, and in some it does not; so, by the Stalnaker–Lewis semantics, it cannot be true that it *would* light—only that it *might*.

Keep thinking of me, and the match, in these same 'iffy' circumstances; and let 'strike' remain vague, indeterminate between the successful and unsuccessful striking styles. Now consider the truth or falsehood of the

[24] Lewis would say that the conditional is simply false; Stalnaker that it is neither true nor false, but indeterminate in truth-value. For comparison of the views (and defenses of Lewis's judgment about such cases), see Lewis, *Counterfactuals*, 77–83; and Jonathan Bennett, *A Philosophical Guide to Conditionals* (Oxford: Clarendon Press, 2003), 183–9.

subjunctive conditional 'If I were to strike the match, it would light' on the supposition that the match actually is struck and, as luck would have it, struck in one of those ways that would cause it to ignite. Intuitions diverge about the general way to ascribe truth-conditions to a subjunctive conditional with true antecedent and consequent. Lewis's official view is that, when antecedent and consequent are both true, a subjunctive conditional is automatically true—although its truth may in that case be due entirely to the truth of the antecedent and consequent, not to any interesting connection between the facts they report. Suppose someone had said, early in 1963, 'If C. S. Lewis were to die on November 22, 1963, then John F. Kennedy would, too.' The conditional is true, (David) Lewis would say, though not because of any important connection between the two events. Its truth is due to two independent facts: C. S. Lewis died on 22 November 1963, and so did J. F. K. Others would say that a conditional like this one is false, and Lewis confesses to feeling a slight tug in that direction. To give in to it (as I am inclined to do) would be to accept the fact that sometimes there are worlds that, although they are different from the actual world, are nevertheless as close to it as it is to itself, for purposes of assessing subjunctives.

Consider what each of these parties will say about the case in which the vague hypothesis (*that I strike the match*) is true, and the match actually ignites, although the match could just as well have been struck in ways that would not have caused it to light: Those who favor Lewis's official account of subjunctives with true antecedent and consequent will say that the conditional about the match is in fact true; although they must admit that its truth depends upon the truth of both antecedent and consequent; those who favor the latter approach should simply deny that it is true, no different than a subjunctive about the striking of this sort of match in these sorts of circumstances when it is contrary-to-fact. As luck would have it, the match was struck and flame resulted, they will say; but it could just as easily have been struck without producing a flame; so 'If I were to strike the match, it would ignite' is false (or at least not true). What is true is merely that, if I were to strike the match, it *might* ignite . . . but, then again, it might not.

The usual way to apply the Stalnaker–Lewis approach to subjunctives and counterfactuals about indeterministic processes treats a hypothetical event that has some chance of causing one outcome, and some chance of

causing another, as relevantly similar to the case of the vague hypothesis *that I strike the match* in the circumstances just described. In both sorts of case, when the antecedent of the conditional is false, some of the nearest worlds that make the antecedent true are ones in which the consequent is true as well; while others, though equally near, make the consequent false. Suppose that there are two worlds in which the imagined striking occurs in exactly the same way in all relevant physical respects; but the actual laws governing situations like this are indeterministic, leaving it up in the air whether the match will ignite. In that case, there would seem to be two worlds, equally 'nearby', containing the striking together with the actual laws, plus as much of the categorical past as can be retained consistently with the supposition that the match is struck; and, in one of these worlds the match lights while in the other it does not. If nearness of world is determined by categorical similarity of past and laws alone, then, in the indeterministic case, neither world is closer to the actual world. And if so, the conditional 'If the match were struck, then it would light' is false. Instead, it is merely true that, if the match were struck, it *might* light . . . but, then again, it might not.

In the case of a subjunctive conditional with true antecedent and consequent describing an indeterministic striking and ignition, the same two alternatives present themselves as in the case of the similar conditional infected by vagueness: either the relevant subjunctive conditional is true, but in a way that depends upon the truth of antecedent and consequent, and not because of a deep connection between the two events they describe; or it is false, because worlds with a striking and no ignition are just as close, for the purposes of assessing such conditionals, as the actual world is to itself.

Now, the libertarian thinks there is indeterminism at work in our choices. Suppose a libertarian accepts the Stalnaker–Lewis account of the truth conditions for counterfactuals and subjunctives about indeterministic situations, as just described. She will have to say that the conditional, 'Were Eve tempted in such-and-such specific ways (in conditions that leave her genuinely free), then she would freely sin', is not true—at least, if Eve is never in fact tempted in this way. A libertarian who favors Lewis's truth-conditions for subjunctives will allow that, if in fact she *is* so tempted, and she *does* freely sin, then the conditional will be true. But it is true, she will say, only because of the truth of both the antecedent and the consequent. Those who favor the other approach will say it is simply false,

since Eve could 'just as easily' have refrained from sinning. In any case, the libertarian who applies the Stalnaker–Lewis semantics in either of the standard ways to subjunctives about indeterministic outcomes will reach the same conclusion: If, at the first stage, God only considers propositions that are true independently of his creative choices, that stage will not contain subjunctives about the outcomes of genuinely free choices. Such conditionals are either not true, or else their truth is dependent upon the truth of their antecedents—and therefore dependent upon God's choice to create free creatures.

The Molinist, however, denies that this is the right way to think about the truth-conditions for subjunctives describing what would happen in indeterministic settings—at least, the ones involving free creatures. There can be 'brute facts' about what would happen if this or that indeterministic situation were to obtain—facts that are not settled by the nearness of worlds, at least if nearness is measured by the categorical facts about the past plus the (indeterministic) laws. According to the Molinist, it is simply a contingent fact that, if Eve were tempted in such-and-such ways (in conditions in which the outcome is left undetermined by the actual laws of nature), she would freely sin. It could have turned out otherwise, but that is how it is. And God knows this contingent fact; it is true independently of any choices God makes, and so is true in some worlds where Eve is never tempted in this way.

Philosophers attracted to the Stalnaker–Lewis semantics, as I have described it, will find it hard to stomach 'brutely true' subjunctives and counterfactuals of this sort. They will draw an obvious moral from the comparison of conditionals about free choices with the conditionals about various match-lighting scenarios: subjunctive conditionals about indeterministic situations—including those describing what free creatures would do—cannot be available for God's use at the first stage. The CFs that will turn out to be contrary-to-fact are simply not true; they are like the contrary-to-fact conditionals about what the match would do under the assumption that the conditions and indeterministic laws do not settle whether it lights. The CFs that happen to have true antecedents and consequents might qualify as true, at some stage or other, though many will doubt even this. Still, their truth depends upon how things happen to go in the future, not upon a reliable linkage between the truth of the antecedent and that of the consequent. So these CFs, if true at all, are not

true independently of God's decision to create; thus, they are not available at the first stage.

Most Molinists still pay lip service to the Stalnaker–Lewis semantics for counterfactuals by insisting that sameness of CFs is one important respect of similarity between worlds. But, at least for indeterministic situations involving agents, they deny that orthodox application of the semantics yields the correct truth-conditions for subjunctives and counterfactuals. For now, I am prepared to grant them this departure from orthodoxy. What is important for the purposes of my argument is that Molinist CFs are 'brutely true'—that is, not grounded in categorical facts about the past, plus the laws of nature.

'Ultima facie' CFs

Suppose the conditions explicitly mentioned in the antecedent of a brutely true CF leave out lots of (seemingly) irrelevant detail about the past. Suppose, for instance, that it is brutely true, before God decides whether to create Eve, that: (i) If Eve were tempted to sin while living in a garden, she would do so. This CF is silent about the size of the garden, the form of the temptation, the time at which the temptation occurs, and so on. Could (i) and the following two CFs all be brutely true together? (ii) If Eve were tempted by a snake in a garden, she would sin, but (iii) if she were tempted by a toad in a garden, she would not. It is not obvious that the conjunction of the three cannot be true. On most accounts of the logic of subjunctives and counterfactuals, they do not obey 'Antecedent Strengthening'. Letting '>' represent the connection between antecedent and consequent in a subjunctive conditional, and 'A', 'B', 'C' stand in for declarative sentences (which are thought of as taking on the subjunctive mood when connected by '>'), the following schematic principle captures this fact about the logic of subjunctive conditionals: From $A > C$ it does not follow that $(A \& B) > C$. Although one is tempted to see some kind of inconsistency between (i) and (iii), the antecedent of (iii) has more content than that of (i), and there is no straightforward problem with their both being true. (i), (ii), and (iii) might be like the trio: (a) If Bush were to resign today, Cheney would become president; (b) if Bush were to resign today and Mick Jagger were to die today, Cheney would become president; (c) if Bush were to resign today and Cheney were to die today, Cheney would not become president.

One objection to (i), (ii), and (iii)'s being true together is that they imply problematic subjunctives constraining God's choices; for (i) and (iii) imply that: (iv) were Eve to be tempted by an animal in a garden, then she would not be tempted by a toad. The logic of subjunctive conditionals seems pretty clearly to support this form of inference: If $A > B$ and $(A \& C) > \sim B$, then $A > \sim C$. But, according to the Molinist, CFs are consulted *before* God has chosen whether to create Eve, and *before* he has chosen whether to tempt her in a garden or on a boat, with a toad or with a goat, etc. And they are supposed to be true independently of God's will. So, with (i), (ii), and (iii) brutely true independently of any resolutions or choices God has made, he knows that were he to put Eve in a garden and allow her to be tempted, he would not choose a toad—at least, he knows this so long as we assume that simple inferences from things God knows at the first stage are also available at that stage.

I set these worries aside, however. The possibility of all three CFs being brutely true would imply that some CFs, although true, are not useful to God. Should God tempt Eve with a talking toad in a garden, in light of the brutal truth of (i), (ii), and (iii)? One might have thought that (i) would be all God needs to know, (ii) and (iii) providing more information then he needs. If he does not want her to sin, (i) counsels against tempting her in a garden by means of anything—snake, toad, goat, angel, etc. But does (i) really provide such counsel? No, not if it is consistent with (iii) being true *and equally relevant to God's choices*. Given that (iii) is brutely true as well, (i) becomes irrelevant to the specific question whether God should tempt her with a toad in a garden. As mentioned earlier, standard accounts of the logic of conditionals imply that (i) and (iii) together imply that, (iv) were she tempted in a garden, it would not be by means of a toad. But the truth of (i) and (iv), and God's desire that Eve not sin, do not provide any kind of reason for God not to tempt her with a talking toad, if (iii) is also true.

If (i), (ii), and (iii) can be brutely true together, the brutal truth of (i) by itself would not be the kind of CF God needs in making his plans. What God needs are CFs that do not switch from true to false when more details are added to their antecedents—and that is what happens to (i), if (iii) is also true. In order to have solid reasons for his creative decisions at the first stage—for instance, the decision not to tempt Eve in a garden *with a toad*, a choice God made 'before' deciding whether to use a toad or snake if he tempts Eve in a garden—God needs true conditionals that can withstand

a certain kind of strengthening of the antecedent without changing truth-value. What kind of strengthening? Not just any old strengthening—for example, strengthenings of the antecedent that make it impossible, or impossible in conjunction with the consequent, are obviously irrelevant, whatever their effect might be on a conditional's truth-value. God need only worry about stronger antecedents if they produce a conditional that is relevant to God's pre-creation deliberation about whether to allow the choice in question. The choice is made at some 'stage' in God's foreknowledge. At any given stage, there are many things 'settled' and many others left 'open', and a wide range of things God could choose to 'settle' on the basis of nothing more than the knowledge in that stage. Presumably, for any set of propositions that could represent a stage in God's foreknowledge, there are a variety of 'complete actions' that God could decide to take on the basis of the knowledge in that stage alone. A complete action, relative to a stage, is a proposition satisfying two conditions: (1) God could decide to 'make it true' on the basis of the knowledge in that stage (to use Plantinga's terminology, it corresponds to a state of affairs God could, at that stage, decide to 'strongly actualize'), and (2) the result of this decision would be a new stage in God's foreknowledge—an *interestingly* new stage, one that contains more information than just what follows from the old stage and the fact that God took this decision. For a genuinely new stage to result, God's action must have left something undetermined that will be 'settled' by events he does not directly bring about. Adding the knowledge of this outcome to the information God used in choosing his action results in a new stage of God's foreknowledge.

The useful CFs might be called '*ultima facie* CFs'. The notion of an *ultima facie* conditional can be defined schematically in terms of 'complete actions relative to a stage':

(UF) Relative to stage S in God's foreknowledge, it is an *ultima facie* truth that $A > B =_{df}$ (1) It is true that $A > B$; and (2) for each complete action x that God could decide to take at stage S, and every proposition p that is known at stage S; if God's deciding to do x is compatible with its being true that A, then the following conditional is true: If God were to do x and p were true and it was the case that A, then it would be that B.

If (i) can be brutely true in conjunction with the brutal truth of (iii), (i) would only be a *prima facie* reason not to put Eve in a garden to face

temptation; it would provide no reason for God not to put her in a garden and tempt her by means of a *toad*. If, however, (i) were an *ultima facie* truth at the first stage, it would provide God with an *ultima facie* reason not to allow Eve to be tempted in a garden, whether by toad, snake, or any other means. Given everything else true at the first stage, and any complete action God could take compatible with Eve's being tempted in a garden, were Eve to be tempted in a garden, she would sin.

As noted earlier, if God could know (i), (ii), and (iii) before he has decided whether to tempt Eve in a garden with a toad or snake, he would be able to know things from which any moderately intelligent person could infer that God will not use a toad in a garden *before* God himself has decided not to use a toad in a garden. Because of the seeming absurdity of this result, I think the Molinist should say that every CF that is true at the first stage is an *ultima facie* conditional; so that, if (i) is brutely true at the first stage, (iii) cannot be, and vice versa. Whatever one thinks about this issue, however, only the *ultima facie* CFs true at the first stage will be of use to God as reasons for his creative choices. Since only *ultima facie* conditionals are of importance, the qualification will usually be dropped.

The Potential Relevance of Distant and Irrelevant Differences

Suppose it is the case that, if I were to strike a certain match right now, then it would light—because all the conditions are right for lighting and the actual laws are deterministic, implying that even the feeblest of scrapes would generate a flame. In that case, one could add all sorts of descriptions of past and future circumstances to the antecedent of this subjunctive without changing the conditional's truth-value—so long as the changes do not have an impact upon what the laws would be like, or what events would occur in the spatio-temporal region of the match just prior to my striking it. If I were to strike the match and Bush were in the White House scratching his nose, then the match would light; if I were to strike it and Bush were in the White House napping, then the match would still light. What Bush does far away should not make a difference to the ignition of the match; if the facts added to the antecedent are sufficiently far-removed from the circumstances at hand, and if the laws of nature would not have to be different for these facts to obtain, then adding them will not affect the truth of the conditional. Why not? Because, in the deterministic case, with laws implying that effects are a function of local causes, the truth

of the conditional is settled by the laws, the nature of the hypothetical striking, and categorical facts involving causally relevant goings-on in the actual world just prior to the time at which the ignition would have taken place. Altering events far away will not make a difference, so long as nearby conditions (the dryness of the match, the presence of oxygen, etc.) and the laws of nature are held constant.

A similar moral follows in the indeterministic case, so long as subjunctive conditionals about indeterministic processes are treated in the orthodox way described earlier. Such conditionals turned out to be false, at least when contrary-to-fact, because worlds in which antecedent is true and consequent is true were no closer to the actual world than ones in which antecedent is true and consequent false. Adding descriptions of distant differences from the actual world, while keeping the laws and events near the indeterministic striking the same, should not help to make these conditionals true—at least, so long as the laws that actually govern the events in question are local in character, as they seem in general to be in the actual world. What is going on at a given place and time seems to depend mainly upon how things are in the vicinity just prior to that time.[25] If so, differences in conditions far away in space and time will not, typically, matter—unless they are differences that would require radically different laws of nature.

The Molinist could not use similar reasoning to show that her CFs are insensitive to trivial differences far away from the events described. The grounds for the truth of a CF are very different than the ones just described; they are not to be found in the vicinity of the effect, even if all causal influences are local. Adding causally irrelevant categorical details about the distant past to the antecedent of a CF may well produce a CF with a different truth-value—or so I shall argue. Here is an example of the sort of thing I want to force the Molinist to allow. Suppose that it is not true (not *ultima facie* true, at any rate) that, if Eve were tempted by a snake in a garden, she would sin; so (ii), above, is false. The following CF could nevertheless be true: (v) If a certain angel sang a certain song one billion years prior to the events in the Garden of Eden, and Eve were tempted by a snake in a garden, then Eve would sin. Adding that little bit of causally irrelevant

[25] 'Collapse theories' of quantum phenomena allow for a kind of action-at-a-distance; and, in general, it must be admitted that there is great controversy about the spatio-temporal boundaries of the conditions immediately causally relevant to certain events.

detail about the angel's song will, in some worlds, make the difference between a false and a true conditional. The reverse is possible, as well: a CF that says very little about the rest of the world might well be *ultima facie*, rendering such differences in distant detail inconsequential (this possibility is not strictly relevant to my argument). (i), for example, could be an *ultima facie* CF; in which case, it matters not what might have happened in the past or what might happen in the future, nor what sort of animal might be used to tempt Eve. So long as the temptation were in a garden, then she would succumb.

CFs should display this sort of elasticity because their truth is brute. The search for *ultima facie* CFs usable at the first stage is not a search for antecedent conditions that would be *sufficient* to *cause* a certain free decision. By hypothesis, the antecedent of a useful CF describes an indeterministic situation, and whatever further conditions are added must leave it so, if the choice described in the consequent remains free. Compare one possible complete history of the universe up to (but not including) the occurrence of a free choice, with another such complete history that differs only in some tiny way in the distant past (for example, a difference in the song an angel sings, or in the swerve of an atom, or in the motions of specks in a space-time that existed before the big bang). Does the truth of a CF with the one complete past as antecedent imply anything about the truth of a CF with the other past for antecedent? If the truth of one of the CFs were grounded in categorical truths about the universe immediately prior to the choice, then distant past events could be relied upon not to make a difference, and either both CFs would be true or neither would be true. But the Molinist's CFs are not grounded in those sorts of local facts, and therefore similarity of local facts will not underwrite a necessary connection between the CFs with antecedents describing histories that differ only distantly. There are possible worlds in which these histories would be followed by one choice; and possible worlds in which they would be followed by another choice; so why not worlds in which the histories differ in the choices to which they would lead?

The Molinist might argue that, as a matter of necessity, the CFs must both be false or both true, because the differences in the antecedents are too distant from and irrelevant to the type of choice in question to make a difference to the outcome. This would be a strong claim: that some particular distant difference between two categorical histories could not

possibly make a difference; or, to put it another way, that the disjunction of the two antecedents is, in every possible world, sufficiently detailed to yield a true *ultima facie* CF about what the agent would do. A Molinist who made this claim would confront the question: What is the distance beyond which one can ignore small differences in the categorical history of a choice? Is it a temporal one? A spatial one? Is the boundary set by some kind of limit on the scale of the differences, relative to the size of the agent? Is it a complicated function of many such factors? I will let the expression 'not differing beyond degree n' do duty for whatever particular limit a Molinist might propose; and make the simplifying assumption that all choices are between just two alternatives. Then, the thesis that there is a limit on the relevance of past conditions to a given situation in which a free choice would occur can be expressed as follows:

(LIMIT) Consider any x, e, w, H, A, and B, such that: e is a free choice between alternative actions A and B on the part of agent x in possible world w, and H is the complete categorical history of w prior to e. There is a set of complete categorical histories, H^*, that includes H and all histories of worlds that differ from it less than degree n, and, in every possible world, one or the other of these is an *ultima facie* CF: (a) if x were to choose between A and B after one of the histories in H^*, x would choose A; or (b) if x were to choose between A and B after one of the histories in H^*, x would choose B.

But where are we to suppose the boundary lies between the prior conditions that are relevant, and the prior conditions that are not relevant? What should take the place of 'n'? The standard Stalnaker–Lewis semantics for subjunctives does not imply LIMIT, for any value of n. Because free choices are (by hypothesis) always the result of indeterministic processes, a complete description of all the causally relevant nearby categorical conditions plus the laws of nature does not settle the truth of a CF concerning a person's free choice. Every enrichment of the antecedent by means of categorical facts about further, more distant events is a different conditional about an indeterministic situation, its truth still ungrounded, according to the orthodox Stalnaker–Lewis account of such conditionals. Could the Molinist suppose that some version of LIMIT is a metaphysical necessity, nevertheless? The problem with doing so is that every choice of a limit on what is relevant seems arbitrary; but arbitrary cut-offs are poor

candidates for metaphysical necessities (to put it mildly); and LIMIT must be necessary, if true at all. It purports to describe the space of possible worlds, and so is not something that could be true in one and false in another.[26]

Where might the Molinist suppose the spatio-temporal limit lies—the border beyond which categorical facts could be added to or subtracted from the antecedent of a CF with no possible danger of producing a CF of differing truth-value? How far away, spatio-temporally or otherwise, can categorical facts be and still 'make a difference', in this sense, to CFs? A Molinist might, I suppose, think that only adding or subtracting categorical information about events within the past lightcone of a point could be relevant to what would have happened at that point. I will grant this much of a restriction, for the sake of argument, at least. Still, if the causal process leading up to an indeterministic event is continuous, there is no such thing as *the* set of immediately preceding causally relevant conditions that fall within the event's past lightcone. Suppose that *R* is the space-time location at which a certain agent makes a free decision. Had everything been the same in regions arbitrarily close to *R* within *R*'s rearward lightcone, the circumstances would still have been perfect for an indeterministic decision of exactly the same general character at *R* (or at *R*'s counterpart, in the somewhat altered situation). Which arbitrarily chosen portion of the rearward lightcone should be regarded as the boundary of 'what matters' to the truth of CFs describing the hypothetical act of choosing that occurs at *R*?

Processes filling relativistic space-time might appear to be continuous, admitting no natural answer to this question; but perhaps the lesson of quantum theory should be taken to be the rejection of such continuity. Would it be easier for the Molinist to insist upon the necessity of LIMIT if there were a minimum length to the causally relevant conditions leading up to any event? Could the set of all events 'one quantum-interval earlier' serve as a non-arbitrary boundary around the past history of an event, a limit before which categorical differences could not possible make a difference to the truth-values of CFs?

One problem with the quantum-unit proposal is that, in order to block LIMIT by appeal to such a boundary, the Molinist would have to suppose

[26] Daniel Fogal has floated the idea that an epistemicist about the vagueness of ordinary language might be in a better position to affirm a version of LIMIT (personal communication). I am skeptical, but unable to pursue the idea here.

that time is *necessarily* quantized—a highly problematic assumption, surely. But, really, it does not matter whether a Molinist tries to use some minimal prior interval in quantized space-time or a (seemingly arbitrary) short interval (say, one minute) in continuous space-time as the boundary of the categorical conditions that can make a difference in CFs. Neither choice of a limit is acceptable.

The reason neither a quantum interval nor a minute nor a second will serve as value for *n* is that they conflict with the repeatability, at least in principle, of the categorical circumstances just prior to a given free choice. However unlikely it might be that an individual find herself in precisely the same categorical conditions more than once, there is nothing absolutely impossible in the supposition that it should happen. If the categorical content of one quantum-unit, or one second, or one minute prior to the conclusion of an indeterministic process is—as a matter of necessity—enough to settle CFs about which outcome would happen; then, in any world where that categorical content occurs again and again, the outcome must be the same every time. But, on the hypothesis that the process leading up to a choice is an indeterministic one, repeated occurrences of exactly the same process *can* lead to different outcomes—and, so, there are possible worlds in which they *do*. It follows that a categorical description of the final minute, second, or quantum-unit leading up to the conclusion of an indeterministic process is not enough to generate a true CF specifying what would happen under those conditions—at least, it is not enough as a matter of *necessity*.

What I am arguing for in this section is that the Molinist's commitment to brutely true CFs requires the falsehood of LIMIT. Consider any categorical description *H* of the past leading up to a circumstance of free choice in some possible world. Denying LIMIT means that, however detailed *H* might be, if the description does not completely specify every categorical aspect of the past, then there are details that can be added to *H* that would, *in some possible worlds*, 'make a difference'. In other words, it is *possible* that a CF with *H* categorically enriched in one way is true, while a CF with *H* categorically enriched in a different way is false. Denying LIMIT does not imply that tiny differences *necessarily* matter; nor even that they *actually* matter. For instance, I suppose that, if the Molinist's general picture is right, there is a world in which only ten minutes matters; that is, for every possible indeterministic choice on the part of every possible creature,

a sufficiently detailed categorical description of the previous ten minutes would be enough to yield a true CF about what the creature would choose. My conclusion is secured so long as there are also worlds in which ten minutes is not enough, though perhaps twenty is; and worlds in which twenty minutes is not enough, though perhaps thirty is; etc.

A Vague Limit?

William Lane Craig has suggested (in conversation) that the Molinist could respond in something like the following way: There are nearby categorical differences that can make a difference to the truth of CFs, and there are distant trivialities that cannot possibly make a difference; but there may be no precise cut-off between differences that can and cannot make a difference. The boundary is simply vague, like the difference between a heap of stones and a few stones that are not big enough to be a heap. In order to be relevant to my claims here, the point of this Molinistic rebuttal would have to be that LIMIT can be true even if it is vague—i.e., even though there is no *precise* limit to the distance at which trivialities can make a difference, no precise value for n. Some values of n will make LIMIT definitely true (there are distances beyond which it is, in every possible world, unnecessary to go); others will yield a version of LIMIT that is definitely false; but some intermediate values of n make LIMIT indeterminate in truth-value.

The reply I am (perhaps wrongly) attributing to Craig implies the following: For some history H, and degree of difference n, there is a categorical history H^* differing from H by no more than n; there is a pair of otherwise identical CFs, with H and H^* for antecedent; and it is indeterminate whether the two CFs have the same truth-value. For it to be indeterminate whether both are true or both are false, at least one of them must itself be indeterminate in truth-value. Where could this indeterminacy come from with respect to the two CFs in question? There are, I take it, two possibilities, given that our attention is restricted to providentially useful CFs: vagueness in the concepts used (by God!) to formulate the conditionals, or 'objective vagueness in the world'—indeterminacy that is not due to imprecision in anyone's concepts or language. But both sources of vagueness render a CF unfit for God's use. The CFs at issue are only *ultima facie* CFs that could serve as the basis for God's creative decisions. God will surely not act upon sloppily formulated or imprecise

information, especially when the imprecision can make a difference to his plans. And if it is *objectively* uncertain whether a given history will lead to a certain outcome, bringing about that history will not give God the sort of risk-free providential control over the outcome of the choice that Molinists attribute to him.

For concreteness, suppose that the only relevant factor is time, and that *n* is exact similarity with respect to the previous 24 hours—i.e., suppose that a complete categorical specification of the universe for 24 hours prior to a choice is enough to yield, in every case, a determinate CF to the effect that, if the previous 24 hours were like *that*, the agent would freely do thus-and-so. And suppose that anything less than a complete categorical description of the previous 23 hours will not insure, in every possible world, that there is a definitely true *ultima facie* CF telling God what the creature would freely do in the circumstances described—replacing *n* with 'exact similarity with respect to the previous 23 hours', for example, turns LIMIT into a falsehood. But there are values of *n* between these two that make LIMIT indeterminate in truth-value—an area of indeterminacy in between complete specifications of 24 and 23 hours, in which it can become indeterminate whether differences in the conditions leading up to a choice are relevant. Perhaps, for example, substituting 'exact similarity of history during 23 hours and 30 minutes prior to the choice' for *n* yields a version of LIMIT that is neither definitely true nor definitely false. If *p* is a complete categorical specification of the world for 23 hours and 30 minutes prior to a choice of Eve's, 'If *p* is true, then Eve freely sins' will be neither true nor false in at least *some* possible worlds.

I fail, however, to see how positing this area of indeterminacy takes away the sting of accepting LIMIT. In this idealized example, God cannot, in every possible world, ignore differences greater than 23 hours prior to a choice. It is still a necessary truth that there is a limit beyond which little differences in the antecedents of CFs cannot possibly make a difference to their providential usefulness. And, as I argued above, no natural basis for a necessary truth of this sort can be found in the case of circumstances leading up to indeterministic events. However much vagueness one posits (e.g., seventeen levels of higher-order vagueness), the CFs God *uses* cannot tolerate *any* amount of vagueness. On the vague-limit hypothesis, there will still be a boundary between, on the one hand, antecedents that describe the world (prior to a choice) so thoroughly that they yield determinately

true *ultima facie* CFs in every possible world; and, on the other hand, antecedents that are not sufficiently detailed to insure determinately true CFs in every possible world. I do not see that the introduction of objective indeterminacy in CFs has made acceptance of such a boundary any easier to swallow; the arbitrariness of any line that could be drawn makes every candidate for *n* a poor choice, and the corresponding version of LIMIT an unlikely candidate for being necessarily true—or even necessarily indeterminate.

IV. The Anti-Molinist Argument

Stage One: The Possibility of 'Divine Voodoo Worlds'

My strategy is to argue as follows: on Molinist principles, there is a possible world in which every possible choice of every possible creature could be controlled by God's fiddling with irrelevant details of the creation far removed from those creatures in space and time. In such worlds, not only *could* God control us, he *would be* controlling us. Every possible creature would be subject to something I will call 'transworld manipulability'. Perhaps the likelihood of such a world being the case is comparable to the likelihood of Plantinga's hypothesis of transworld depravity (or transworld sanctity). But if transworld depravity is possible, so is transworld manipulability; and the mere fact that Molinism implies the possibility of transworld manipulability is, I claim, bad enough.

I shall make some simplifying assumptions about what possible worlds with free creatures are like, trusting that the simplifications will make no difference to the argument. I shall assume that every possible world with free creatures has a 'first family'—the group of creatures who first exercise freedom. I shall pretend that there are no possible worlds with infinite pasts in which, for any given time, infinitely many free choices have already been made before that time. I shall also pretend that free choices are always between two alternatives. On these assumptions, when God considers whether to create free creatures, he will be running through infinitely many 'first families', beings that can be created together. Assuming, with Plantinga, that there are individual essences for merely possible creatures, each first family is represented by a set of essences. (A way to formulate

CFs about 'merely possible creatures' without recourse to Plantinga's haecceitistic essences is mentioned in n. 15, above.)

For every possible first family, the Molinist posits CFs about their first choices. A CF about the first choice of a creature need not specify the entire past history of the universe in which the choice is made; in some possible worlds, there are *ultima facie* CFs about the first choices that merely describe circumstances obtaining during the seconds leading up to the choice. But, because LIMIT is false, there are others where the world's history long before the choice matters a great deal; for CFs with antecedents that differ only minutely in their descriptions of much earlier events can imply that different choices are made. I shall assume that, by specifying the 'complete categorical past' leading up to a choice in the antecedent of a CF about that choice, one could eventually arrive at a true CF about what choice would be made. I assume this because, if events causally downstream from the choice are allowed to be among the factors relevant to the truth of CFs about the choice, and some of these include facts that depend upon the character of the choice that is made, the resulting CF might well be of limited providential usefulness. For example, suppose that neither of the following conditionals is an *ultima facie* CF that God can use at the first stage:

(A) If Eve were offered an apple at a time *t* in a world with a complete categorical past of type *H*, then she would freely accept it; and
(B) If Eve were offered an apple at a time *t* in a world with a complete categorical past of type *H*, then she would freely reject it.

And suppose that the only way to enrich the contents of one of these antecedents in such a way that the resulting CF is true would be by adding information about whether Eve is eating an apple shortly after *t*. Could God make use of this kind of CF:

(C) If Eve were offered an apple at *t* in a world with a past of type *H* and were eating an apple shortly after *t* as a result of this choice, then she would freely accept the apple at *t*?

Presumably not, since God could not insure that the antecedent is true except by insuring that Eve chooses to eat the apple—and that he cannot do, in the absence of other CFs by means of which to control this choice.

Could the following CF be providentially useful to God, if it were the only CF relevant to the control of Eve's choice at *t*:

(C) If Eve were offered an apple at *t* in a world with a past of type *H* and she were found chewing on an apple shortly after *t*, then she would freely accept the apple at *t*?

Since God cannot, given libertarian scruples, cause her chewing by causing her (free) choice, in order to use (C) he must somehow insure that, whether or not Eve were to choose to eat the apple, she would still be chewing an apple shortly after *t*. This would limit God's ability to respond to Eve's choice—no matter what she does, he has either to allow or to force her to eat an apple. So, from the Molinist's point of view, it would be best if God could control every possible free choice by means of CFs with antecedents describing just the categorical past relative to the choice. Unsurprisingly, Molinists assume that God has the CFs to do this.[27]

The complete categorical history of the universe prior to the first free choices is what I shall call an 'initial world-type'. For any first family with members $x, y, z, \ldots$, and any initial world-type A in which $x, y, z, \ldots$ could coexist and be the first family, there will be a series of CFs with the occurrence of A for antecedent, and the free choices that would be made by $x, y, z, \ldots$, for their consequents. Since the initial world type A leaves each of their choices indeterministically 'open', and God does not cause them to do one thing rather than another, there must be possible worlds in which the members make every possible combination of choices. So, in those worlds at any rate, there are true CFs affirming that $x, y, z, \ldots$, would make that combination of choices in an A-world. If those CFs are not *actually* true, a world in which A occurs and they make that combination of choices will not be one that God could bring about; it is not a 'feasible' world (to use Thomas Flint's terminology). Still, the possible world is 'out there', and so the CFs that God would have known in such a world must be possible.

Suppose that every first family contains a finite number of individuals, *n*, each one of whom will face an initial undetermined choice between two options. That requires, for each initial world-type A, 2^n possible combinations of CFs describing what everyone would choose in A. Let us

[27] See, e.g., Flint, *Divine Providence*, 47; and Freddoso, Introduction, 50.

suppose that, in fact, our first family included just two people—call them 'Adam' and 'Eve'—and that their first free choices were whether to accept and eat an apple, or to refrain from doing so. To simplify things further, suppose that these choices were made simultaneously and independently. In that case, for each initial world-type *A*, there are only four possible combinations of CFs specifying what the pair would do.

(a) A > Adam accepts, A > Eve accepts
(b) A > Adam refrains, A > Eve refrains
(c) A > Adam refrains, A > Eve accepts
(d) A > Adam accepts, A > Eve refrains

Suppose (a) is the case, and that refraining is the blameless choice. Suppose, too, that *A* includes many objects far away from Adam and Eve, the disposition of which is relatively inconsequential to God's purposes; and that there are three trivial changes in these distant things yielding initial world-types *A*1, *A*2, *A*3—circumstances in which Adam and Eve would also have existed and have faced the same choices. This generates further possible combinations of CFs:

(a1) A1 > Adam accepts, A1 > Eve accepts
(b1) A1 > Adam refrains, A1 > Eve refrains
(c1) A1 > Adam refrains, A1 > Eve accepts
(d1) A1 > Adam accepts, A1 > Eve refrains

(a2) A2 > Adam accepts, A2 > Eve accepts
(b2) A2 > Adam refrains, A2 > Eve refrains
(c2) A2 > Adam refrains, A2 > Eve accepts
(d2) A2 > Adam accepts, A2 > Eve refrains

(a3) A3 > Adam accepts, A3 > Eve accepts
(b3) A3 > Adam refrains, A3 > Eve refrains
(c3) A3 > Adam refrains, A3 > Eve accepts
(d3) A3 > Adam accepts, A3 > Eve refrains

Because LIMIT is false, there are possible worlds in which these little changes in the initial world-types generate differences in the CFs specifying what Adam and Eve would do. So, there is a world in which *A* > Adam accepts, but *A1* > Adam refrains; and one in which *A* > Adam refrains and *A1* > Adam accepts. And likewise for Eve. Furthermore, since their choices are made independently, and the CFs describing what they would

do are brute facts, each combination of CFs about Adam and Eve should be possible, as well—every combination that results from taking one pair of CFs with *A* for antecedent, another pair of CFs with *A1* for antecedent, etc.

The fact that these CFs are brutely true is relevant to the plausibility of the recombination principle I have just affirmed. If they were subjunctives grounded in categorical facts about similarity of worlds, one might think that, if the difference between *A* and *A1* is enough to make a difference to the question whether Adam would sin, then it might well be the sort of difference that would require a change in what Eve would do as well. If evacuating the air in a room would make a difference to whether one match would light, it ought to make a difference to whether another, similar match would light in that same room. Since the categorical facts described in *A* (together with categorical facts about the actual world) are not sufficient to determine which CFs are true, there is no reason to expect that the CFs about Adam and Eve should be linked in such a way that certain combinations are ruled out.

We can easily test whether contemporary Molinists ought to accept my principle of recombination: Are they prepared to make use of Alvin Plantinga's Molinistic version of the Free Will Defense? If so, then they have no right to balk at my assertion that, given Molinism, all these combinations are possible. In his Free Will Defense, Plantinga makes a certain claim about the way CFs could have turned out for every possible free creature.[28] Pretend that Adam and Eve are the only possible people, that *A*, *A1*, *A2*, and *A3* are the only possible initial world states compatible with their existence, and that their first choices are the only important ones they could make. Given these radical simplifying assumptions, Plantinga's claim amounts to the insistence that the following combination of CFs is a possible one:

TD: (a), (a1), (a2), (a3) (Transworld depravity)

Assuming Adam and Eve exhaust the possible free creatures, if the CFs had turned out to be TD, then, if God wanted to create free creatures at all, he would have to create free creatures each of whom sinned. Plantinga uses the possibility of TD to show that, given libertarianism about freedom, it is possible that there be no way for God to insure that everyone always freely

[28] Plantinga, *The Nature of Necessity* (Oxford: Clarendon Press), 169–90.

does what is right—even a God who exercises absolute providential control over his creation by Molinistic means. Now, an opponent of Plantinga's Free Will Defense might claim that TD is not a possible combination of CFs. The *truly* unreasonable opponent might even propose that, although TD is not possible, the following combination *is* possible:

TS: (b), (b1), (b2), (b3) (Transworld sanctity)

So, according to this unreasonable opponent, God could not possibly have been stuck in the extreme transworld depravity scenario, with CFs implying that every possible individual sins upon every opportunity; but he could have found himself with CFs implying that they always do the right thing. It would only be slightly more reasonable to claim that neither TD nor TS is a possible combination, though all the intermediate combinations are possible. But neither response to Plantinga seems reasonable. Given the assumption of Molinism, and the bruteness of its CFs, Plantinga is right to suppose that every combination, including TD, is a genuine possibility.[29]

Of course most of the combinations will be a mixed bag. Most, unlike TD and TS, allow God to choose initial states in which Adam and Eve would make different choices, and different choices from one another. Some will allow God to decide whether to have Adam or Eve sin, but not allow him to create a world in which both sin or neither sins. But the following sort of world would allow God complete control over the way Adam and Eve behave:

TM: (a), (b1), (c2), (d3) (Transworld manipulability)

In this case, each possible combination of free choices open to Adam and Eve can be selected by God. All he need do is fiddle with the tiny, distant differences between *A*, *A1*, *A2*, and *A3*. The worlds in which all essences display transworld manipulability might be called 'Divine Voodoo' worlds, because the CFs that happen to be true provide God with the analogue

[29] For an argument that, given certain assumptions about the number of possible free creatures, and the numbers of choices they could make, the probability of transworld depravity will be infinitesimal, see Josh Rasmussen, 'On Creating Worlds Without Evil—Given Divine Counterfactual Knowledge', *Religious Studies*, 40 (2004), 457–70. If his arguments go through, then the possibility I call 'divine voodoo' will be equally unlikely. This is less problematic for me than for many Free Will Defenders, however. I have no stake in whether all possible free creatures are *actually* transworld manipulable; but some users of Plantinga's Free Will Defense (though not Plantinga himself) are committed to the actual transworld depravity of all possible free creatures, or at least the transworld depravity of any sizable group of them that could coexist and exercise significant freedom.

of a set of voodoo dolls or a remote control device—whatever he wants Adam and Eve or any other creature to do, he can insure that they do it by manipulating insignificant details of the creation far away from the creatures themselves. For the moment, I shall continue to pretend that Adam and Eve represent all the possible free beings, and that there are only four initial world-states compatible with the existence of free creatures. In that case, if TM were true, God would have absolute control over every free creature he could make, and the control would be exercised by 'pushing' spatio-temporally distant 'buttons'. So far, the CFs under consideration only give God voodoo control over the *initial* decisions of these creatures. But, given the possibility of enough tiny differences far away, there is the possibility of further initial world-types allowing for divine control over the outcomes of all circumstances of free choice that could possibly develop in worlds that begin with Adam and Eve.

What happens when we relax some of the absurd simplifications in this picture—for instance, the assumption that there are only four possible initial world-types compatible with free creatures (an assumption we had better relax, given the possibility of further choices which must be correlated with different 'buttons')? What is crucial about the four world-types in the toy 'Adam-and-Eve' example, is that they differ only in tiny ways, and that these differences are not of major importance to God's plans—so it costs God nothing to choose *A* rather than *A1*, *A2*, or *A3*, if *A* is required to insure that Adam and Eve do as he wishes. One might worry that my claims about Adam and Eve cannot be generalized to cover all possible creatures. Consider free creatures that could be created in worlds with initial world-types that *preclude* all tiny, distant differences that do not matter to God. Would they be at least partially immune to God's control in *every* world?

Here is the sort of first family that might be thought to be immune, in some of the situations in which it could be created, to control by means of distant and irrelevant factors. Consider a first family consisting of two angels, each a simple substance, preceded by nothing whatsoever and accompanied by nothing whatsoever. As the pair of first created beings come into existence, they are given their choice of two songs to sing. There might seem to be a big difference between the initial world-type that precedes their creation (one might call it the 'null-type', a past history consisting of nothing at all) and any initial world-type containing tiny,

distant things that could be counterfactually relevant to the choices of the angels. And so one might conclude that, inevitably, the only kinds of 'buttons' that might be available to control the angels' choice of songs would involve creating a radically different world—in which case it is harder to interpret the 'pushing' of the 'buttons' as tiny changes about which God would be indifferent.

I am suspicious of this line of defense, however. Even first families in a null-type world might well be subject to divine voodoo, if they are creatures that could have existed and been faced with the same kind of choice in a world where God first (or simultaneously) created a causally unrelated universe with its own contents. I suppose some space-time manifolds, though devoid of free creatures, are worth creating in their own right, for aesthetic reasons. God seems to have seen value in the creation of a lifeless, immensely complicated, evolving universe—namely, our own, throughout the vast majority of its history. Among the perfectly good universes compatible with the angels as first family, then, I imagine that some contain earlier space-times filled with, for example, glowing, swirling gases—perhaps something like spiral nebulae, for example, some of which were apparently worth *actually* creating even though they have little if anything to do with free creatures (so far as we know!). The gaseous universe passes away, and has no effect on the angels. But the contents of this earlier universe could have been created in infinitely many distinct, equally lovely configurations, evolving according to one of infinitely many laws of development. Some such prior universes could be very small, differing only minutely from a world with the null-type beginning; others could be astonishingly complex. Given Molinism, the falsehood of LIMIT, and Plantinga-style recombination principles, the following combination of CFs is possible: were the null-type world created, the angels would both choose song number one; were a gaseous world of one sort to have preceded them instead, they would have chosen song number two; were a slightly different sort of gaseous world to have preceded them, one would have chosen song one, the other song two; etc. In that case, even if God creates them in the null-type world, there are little differences in equally lovely worlds God could have created; and, in some possible worlds, the CFs with these alternative initial world histories in their antecedents give him complete control over the angels' choices. By deciding whether or not to create some other, aesthetically pleasing things, God decides what

songs the angels will sing, and he is free to choose any possible combination he likes.

Stage Two: Why Even the Possibility of Divine Voodoo Worlds is Problematic

Stage one of my argument can be summed up as follows: The Molinist posits truths about what every possible free creature would do under every possible indeterministic circumstance of choice. On Molinistic principles, there must be some possible worlds in which little changes in trivial features of the distant past (or in causally unconnected regions) can be used by God to control the choices that would be made by his creatures. So there was a chance, however small (though perhaps no smaller than the chances of transworld depravity), that God had found himself confronted by essences each of which displays transworld manipulability. In that case, if he creates free creatures at all, then he creates creatures over which he has absolute voodoo-like control—control exercised by his determination of distant, relatively trivial details about inanimate parts of the universe.

Some of my best friends are Molinists; and many of them are prepared to accept everything for which I have argued so far: it is possible for distant differences to make a difference in CFs, and such CFs could have been combined in ways that would have given God absolute control over all possible creatures. But they see no problem in countenancing this—as an extremely unlikely possibility. The existence of CFs, and their use by God at the first stage in his providential planning of a world, does not render a person unfree, say these Molinists—and the true CFs would not do so, even in the unlikely event that they gave God voodoo-like powers over every possible free creature. So long as God is not able to pick and choose which CFs are true, we remain free, even if the true CFs imply that we can be controlled by the pushing of buttons or twiddling of knobs far away in space and time.

But how could I still be free if someone else possessed the means to determine the outcome of every possible kind of choice with which I might be confronted? (By 'determine' I mean simply 'decide what it will be'—a perfectly good sense of the term.) Perhaps if the person who possessed the means refused to make use of it, I might remain free; but that is not an option for God, if he creates someone who is transworld-manipulable. Consider the angels, whose choices are counterfactually linked to the prior presence or absence of various aesthetically pleasing patterns in a swirling

cosmic dust. Even if God creates the angels in the null-type world, he does not fail to control their choice of song merely because he did not *actually* create the patterns that would have led to their choosing differently. 'Not pushing any button' on a remote control can be a way of controlling the things with which the remote is counterfactually linked. Suppose that, if I were to push one button, the TV would turn off; if I were to push another, the TV would switch channels; but if I were to refrain from pushing buttons altogether, the TV would explode, killing everyone in the room. With the remote in my hand, and full knowledge of these conditionals, I cannot claim that deliberate refusal to push a button was a case of having control but not using it. Given God's knowledge, in advance, of all the CFs about the angels, his choice of a null-type history must count as deliberate control of the angels' choice of song, so long as alternative histories involving lovely patterns of dust were counterfactually linked to the total range of the angels' choices.

Pursuit of the remote control analogy makes clear just how difficult it is to believe that transworld-manipulable creatures would be free. Which CFs are true is, according to the Molinist, not up to God—they are 'given' to God by . . . reality, or contingency, or 'the way things just happen to be'. In every world, they provide God with something like a remote control device; that is how the Molinist explains God's risk-free providential control. But in some worlds the device is nearly useless. It is as though there were a remote control manufacturer—an 'independent contractor' over whom God has no authority, operating 'before' God decides what to create—and the quality of the remote control produced by this manufacturer varies from one possible world to another. When God gets the manufacturer's handbook (i.e., the list of true CFs), explaining what the buttons do, he may find that facts about parts of the physical world far from his free creatures would be relevant to controlling them, providentially. But he might also find that, no matter what the world is like beforehand, any free creature offered an apple, say, would accept it. The buttons that, in some worlds, control apple-choosings, have been disconnected. Similarly, in worlds where transworld depravity runs utterly rampant, it turns out that, no matter what buttons might be pushed before creatures face morally significant decisions, they would sin.

The chanciness of God's being given a really good remote does not seem to me to be at all relevant to the degree of freedom possessed by the

creatures he controls, once he has one in hand. Suppose a scientist (mad, as usual) makes a super-sophisticated humanoid robot, the behavior of which is designed to be somewhat indeterministic. She then makes a device that looks for all the world like a remote control. The remote is not hooked up to the robot by wires, or radio waves, or any known method of information transfer. It is also an indeterministic matter, at the time at which the remote is made, whether there will be any correlations between pushing buttons on the remote and the behavior of the robot. But if she is really lucky, the device and the robot will 'magically' link up, so that, despite the apparent indeterminacy within the robot (indeterminacy understood as latitude left by the laws of nature), she can get it to do whatever she wants among all the physically possible options open to it at any given time, just by pushing buttons on the remote. And, somehow, she knows (with certainty) whether such a link has been established; and she knows (with certainty) which combinations of button-pushings (and failures to push buttons) are correlated with which robotic actions (and failures to act). If she is lucky, she will have absolute control over the robot—even though there was nothing she could have done ahead of time to insure that she had such control. But, given that the counterfactual link has been established, she can decide exactly what to have the robot do, in every circumstance. To make the case even more like that of the divine voodoo in worlds where everyone suffers from transworld manipulability, another modification is needed: let us suppose that she has just one chance to make such a device. Only the first remote created in the proximity of a given robot has any chance of linking up with it in this way; if it fails, or only allows for partial control, that is the end of the story; there is no point in trying again.

The link between remote and robot in the story is not exactly causal—or, at least, it is as non-causal as God's control of us by means of CFs. Pushing the buttons does not cause the robot's motions, if 'causing' means something like 'bringing about by means of a transfer of energy'. Still, despite the absence of normal sorts of causal transactions between the remote and the robot, the mad scientist has complete 'counterfactual control' over it. Not even completely refraining from pushing buttons will prevent her from exercising her control, given her knowledge of what the robot will do in every circumstance. If she holds the remote in her hand, and knows what situation the robot is in, and decides not to press any buttons,

she is just as thoroughly in control of what it does as when she pushes buttons.[30]

Adding more robots might limit the mad scientist's control in a way that parallels limitations God might face. Suppose that, if a single remote device were made in the vicinity of an army of robots, there would be a chance of its buttons controlling them all. If the pushing of combinations of buttons were counterfactually linked to each robot's behavior (perhaps different types of behavior would be exhibited by different robots were the scientist to push a certain series of buttons), then her ability to control the army might be very limited. Suppose, for example, that the robots are 'Sentinels' (gigantic, mutant-hunting robots), and that she wants them to kill all mutants. Fortunately (from the perspective of the mutants), for any button she pushes that will set some of them to work hunting down mutants, it puts others to work protecting mutants. Still, if she is extremely lucky, and the device has enough combinations of buttons, the Sentinels could turn out to be completely under her control—for the remote might link up with the robots in such a way that, for every possible combination of actions the robots could take at a given time, she has available to her a way to push buttons that will guarantee that they perform that combination of actions. Then, if she wants them all to hunt mutants they will; and if the mutant-loving Dr Xavier gets hold of the remote, he can insure that they only protect mutants. If the perverse members of the Hellfire club acquire the remote, they will no doubt decide to have the Sentinels do some combination of the two. But whoever controls the ideal remote has complete latitude in choosing what the robots will do; any possible combination of activities can be selected.

Suppose a remote has turned out perfectly, giving the possessor complete control over some creature's every decision and action; and suppose our mad scientist is holding the remote and staring at the creature, vividly aware of the precise ways in which the creature's behavior depends upon her pushing or refraining from pushing certain buttons. Could we possibly

[30] Alexander Pruss has suggested to me that, so long as the person holding the remote control does not *care* about the outcome, failure to push buttons need not count as control, even in the face of vivid knowledge of what would happen were one to push them and what would happen if one did not. But, even if this is true, it is irrelevant to the case of control over our free choices between good and evil options; God is supposed to care a great deal about such things.

regard the subsequent actions of the creature as *freely chosen*? Can someone be under such complete control, and yet remain free?

One thing is certain: If *I* discovered that someone had this sort of control over my decisions, I would conclude that I was not a free agent. And I suspect that most people would have similar reactions. If we would be right to feel this way, what follows for Molinism? Granting the possibility of the CFs turning out in such a way that free creatures are impossible, the Molinist must admit that God *could* have found himself unable to create free creatures at all. The cost of this admission will be explored in the penultimate section. First, I shall try to justify my hypothetical reaction to the discovery that I am transworld manipulable.

Transworld Manipulability and Freedom

Suppose, for concreteness, that, long ago, there was a patch of cosmic dust blowing around in a complex pattern; and that God's selection of a pattern for this dust enabled him to control every choice I could possibly make. Imagine a continuous space-time with precisely located dust particles; and suppose that the (literally) infinitely many precise ways the dust could swirl provide the means to insure that I pick any one of the range of options that would be open to me in any indeterministic situation I could encounter.

Learning that God used the dust to control my every choice would convince me I am not free. But suppose I reach this conclusion by tacitly reasoning as follows: the state of the dust, plus the CFs available to God before creation, would not be within my power to change; but, if the concrete past (also now beyond my power), plus these CFs, together entail that I do something, then I do not act freely. This argument is based on the same principles as van Inwagen's 'Consequence Argument'[31]—the *ur*-argument for incompatibilism, in the minds of many contemporary libertarians—and, if it works when the CFs deliver transworld manipulability, it ought to work when they yield transworld depravity, transworld sanctity, or some more mixed outcome.[32] If my hypothetical reaction ('I'm not free!') to news of my transworld manipulability is only justified by the soundness of this argument, then the possibility of divine voodoo worlds

[31] Cf. Peter van Inwagen, *An Essay on Free Will* (Oxford: Oxford University Press, 1986), ch. 3.

[32] William Hasker has developed several impressive arguments along these lines. Quite different versions may be found in Hasker, *God, Time, and Knowledge*; and id., 'A New Anti-Molinist Argument'.

is playing no special role—except to make vivid just how radically 'up to God' it could be which alternatives are ruled out. Molinists generally respond to this sort of argument by claiming that, in the sense of 'within my power' that is relevant to freedom, I *do* have power over the true CFs: had I chosen otherwise, the CFs would have been different. I have a kind of 'counterfactual control' over the CFs that renders them innocuous.[33]

But even granting the validity of the Molinists' response (at least for the sake of argument), there seems to me to be something *more* going wrong when God exercises divine voodoo power over his creatures. Libertarians ought to think that more is required for a kind of freedom-worth-having than mere indeterminism in the process of choosing. One of these requirements is *not being completely under the control of another person*. Normally, one should have thought that the only ways to control another's choices involve ruling out alternative possibilities. But, if one accepts the Molinist scheme, that turns out to be false.

I tried to avoid overtly *causal* language in describing the counterfactual connections that give the mad scientist control over her creations, or God control over me by means of the dust. But could it be that the word 'control' is already loading the deck against the Molinist? No. It should be clear, from the earlier discussion of the motivations for Molinism, that first-stage knowledge of CFs is important precisely because it seems necessary in order for God to exercise risk-free providential control over all of history. His inability to also determine what the CFs will be provides a bit of a buffer between God's will and our choices; but it does not prevent him from controlling the course of events by means of his advance knowledge of what we will (and would) freely do. The connections between the mad scientist's remote and the robot hold because of the same kinds of conditionals; if God's providential arranging of things so that they turn out as he plans counts as a kind of control over history—and this Molinists are quick to affirm—then the scientist's ability to choose what her creation will do amounts to complete control over its actions. And likewise for God's ability to choose what I do by means of the swirling dust.

The less control I have over someone, the more autonomous the person is. Control, and autonomy, come in degrees. Suppose I suffer from extreme transworld *depravity*, rather than manipulability. In that case, so long as God

[33] Cf. Flint, 'A New Anti-Anti-Molinist Argument'.

has reason to allow me to exist at all, he cannot be held responsible for my choosing evil rather than good; my choosing evil is beyond his control. Likewise, if I am transworld sanctified, he gets no credit for not having created me in situations in which I would have chosen badly. In either case, I display a kind of autonomy, relative to God's goals. If, however, I am transworld manipulable, I present God with no obstacles; he has absolute freedom to determine whether I choose good or evil in every possible situation. Since, by hypothesis, the CFs *in fact* allow God to control me by his disposition of the swirling dust, I am *in fact* under his control—even if the Molinist is right in supposing that I have the 'counterfactual power' to make it the case that I *not* have been controllable by means of the dust. Since the CFs are *true*, I never in fact exercise this power, and so I remain under his control for my entire life. If someone else has the power to decide exactly what to have me do in every circumstance, then, even if I have counterfactual control over whether the person has this power, so long as I do not in fact rob him of this power, I am not free.

The Molinist response I have most often heard to my argument is simply to 'stare me down' at this point. I can be under the complete control of another person—that is, I can place no limits upon his freedom to decide what I will decide—and yet I can be perfectly free (even in the libertarian's robust sense of the term, as opposed to some watered-down, compatibilistic surrogate for 'free'). I am tempted to respond to this claim by alleging that being free analytically entails *not* being under the complete control of another. I do not see how to prove this. But I can at least prove that it is not the 'ruling out of alternative possibilities in advance' that is driving my judgments about these cases. Consider the mad scientist again: suppose that the button-pressing by means of which she is able to control the actions of a robot or a living creature must occur at the exact same moment as the choice—the CFs that give the remote its power are of the form, 'If such-and-such buttons were pressed at *t*, the creature would freely decide, at *t*, to do so-and-so'. The mad scientist can be an agent with libertarian freedom, undecided, right up until the moment of the choice, about what to have the robot or creature do. In that case, alternative possibilities are not ruled out in advance; but my reaction to the case is the same. If the remote affords her *complete* control—the ability to select any choice that is left open to the robot or creature by the laws of nature—then the robot or creature could not be free. Or so it seems to me.

Could God Have Been Unable to Make Free Creatures?

I can see a way for a Molinist to accept my two claims—that transworld manipulability is at least possible, and that a person so completely manipulable would not be free—while nevertheless insisting that, since not every essence has turned out to be transworld manipulable, God was able to create free creatures while knowing exactly what they would freely do under every possible circumstance. Freedom is built into CFs, as I have been using the term; they are conditionals about what choices would freely be taken by creatures in indeterministic circumstances. But, for every CF of the form, 'Were *x* in such-and-such (indeterministic) circumstances, *x* would *freely* do *A*', there is a weaker conditional of indeterministic behavior, 'Were *x* in such-and-such (indeterministic) circumstances, *x* would do *A*'. The Molinist might admit that, if all possible free creatures had turned out to be transworld-manipulable, then God would not have been able to create any of them *as free creatures*. (They still would be *possibly free*. Since their transworld manipulability is not essential to them, there are possible worlds in which they are not under God's complete control—that is, worlds in which the CFs that happen to be true do not give God a very powerful remote control device.) My hypothetical Molinist objector says: Why should the Molinist care much about this distant possibility? What is important is that God know plenty of the weaker sort of conditionals, the ones describing what all possible creatures would do under indeterministic conditions. So long as he knows those kinds of conditionals, and they do not give him extreme manipulative powers, then he can create free creatures while knowing what everyone would do under all possible circumstances.

The Molinist who adopts this strategy would grant the possibility of transworld manipulability for every essence, while admitting its incompatibility with God's creating free creatures. Possible creatures that display extreme forms of transworld depravity or sanctity greatly limit God's options—e.g., creatures who, for every range of actions open to them, would always take the worst option *no matter what*, could not be controlled by fiddling with the conditions, distant or near, leading up to their choices. If God's knowledge of the CFs about me is compatible with my freedom (so long as they do not render me transworld manipulable), then I am extremely free in worlds where I turn out to be transworld depraved or

sanctified, and less free to the extent that the CFs give God more control over my choices, providing him with better and better 'remote controls'. This response requires that freedom come in degrees; but that does not seem so hard to swallow.[34] Our Molinist can agree, then, that God *could* have found himself without the option to create free creatures; but also insist that this possibility was vanishingly small, and so not worth worrying about.[35] Molinism has an exotic possibility as a corollary; but this possibility is really no more exotic or surprising than transworld depravity or sanctity; once a Molinist has recognized these latter two possibilities, is the possibility of transworld manipulability really so hard to accept?[36]

There is a downside for the Molinist who would take this line. She must admit that God only contingently has the power to create genuinely free creatures. Accepting this conclusion strikes me as only marginally better than the first Molinist response I considered: i.e., simply 'staring down' the stories about robots and control by cosmic dust, and claiming that the intentional use of a perfect remote control device is no threat to the freedom of the creatures it is used to control. Here is the challenge for the Molinist who would use the second strategy instead: Does she really want to say that God could have 'woken up' to find that he has been given, by nobody in particular, a remote control device so powerful that, whatever he does, his possession of it prevents him from being able to create free creatures? Admitting that God could find himself in this depressing position may not violate the letter of omnipotence, once the notion is formulated with sufficient care.[37] Accepting this conclusion underscores the fact that the Molinist's God, though he does not take risks, is nevertheless *subject to* risk. There was a chance that God's desire to create free creatures would

[34] In correspondence and unpublished work, Daniel Fogal has pointed out to me that freedom's coming in degrees in this way might be vulnerable to the following sort of argument: there would have to be a line between creatures too manipulable to be free *at all* and creatures very manipulable but still free *to some small extent*; but (one might suppose) it is implausible to suppose there is such a line; so either creatures are free no matter how manipulable, or they are not free at all (if Molinism is true). Fogal also suggests that the 'no arbitrary cut-offs' principle at work in this argument is close to the kind of reasoning I make use of in my arguments against LIMIT; and so if I accept the one I should accept the other. I do not think the two appeals to 'no arbitrary cut-offs' stand or fall together, but will set this interesting question aside here.

[35] The Molinist could also use Rasmussen's reasoning (in 'On Creating Worlds Without Evil') to defend the idea that there is zero probability that we are in a divine voodoo world, and argue that this renders its bare possibility innocuous.

[36] Daniel Fogal, Sam Newlands, Mike Rea, David Hunt, and others have pressed me on this point.

[37] As in, for example, Thomas Flint and Alfred J. Freddoso, 'Maximal Power', in Freddoso (ed.), *The Existence and Nature of God* (Notre Dame, Ind.: University of Notre Dame Press, 1983), 81–113.

not be realizable. Fortunately, things seem not to have turned out that way; but they could have.[38]

V. Conclusion

My own judgment is that neither response is plausible. Something has gone terribly wrong if one is forced to admit the possibility of divine voodoo worlds; and no one will be surprised to learn what I think it is: the supposition, dubious to begin with, that free creatures can be infallibly manipulated while remaining free—that they can be deliberately put in circumstances where they freely do something, even though the one who put them in those circumstances has, in advance, infallible knowledge of what they will do. It is the hypothesis of the availability of CFs at the first stage in God's foreknowledge, together with the contingency of the CFs, that has generated the voodoo worlds in which we are too easy to control to be free.

Control by means of Molinist CFs might *seem* consistent with freedom, so long as we do not think about the case in which God would have absolute voodoo control—so long, that is, as we ignore possible worlds like the ones in which God is able to make each of us do any one of the range of choices open to us whenever we face a free decision, by means of a careful choice of a pattern for some swirling dust somewhere in the distant past. But a Molinistic theory of providence requires either conditionals of freedom, or at least conditionals of indeterministic behavior, that do, in some possible worlds, give God this sort of extreme voodoo control. So Molinism requires either the possibility of free action on the part of creatures who, it seems to me, could not really be free; or the possibility of God's being unable to create free creatures at all. Neither alternative is a happy one.

If, as Molinists sometimes allege,[39] their view provides the only way for a libertarian to consistently affirm that God has complete foreknowledge (or complete timeless knowledge) while exercising risk-free providential

[38] It has been pointed out to me that Molinists might already be forced to admit that God could find himself in this situation; for it is tempting to say that, if every possible free creature were afflicted by extreme transworld depravity, a God who could do no wrong would not be able to create any of them.

[39] e.g., Flint, *Divine Providence*, ch. 3.

control; then libertarians should accept at least a part of the Open Theists' controversial package—namely, the thesis that God had to choose whether to create free creatures 'before' knowing what they would do.

VI. Appendix: CFs as Indicative Conditionals

Could Molinists construe CFs as indicative conditionals, rather than counterfactuals or subjunctives? And, if they could, would it affect my argument?

The first question is complicated by the fact that philosophers of language differ radically in their views about the nature of indicative conditionals. As I see it, there are three deep faultlines separating rival theories of indicatives. (1) Some philosophers deny that the assertion of an indicative conditional is typically used to express a proposition; they say utterances of indicative conditionals are not the kinds of speech acts that can properly be evaluated for truth and falsity. (2) Others think indicatives are really material conditionals in disguise. (3) Still others provide truth-conditions that in one way or another incorporate facts about the speaker's epistemic situation. In this appendix, I provide a rough sketch of each approach to indicative conditionals. In each case, it appears that Molinists either could not construe CFs as indicatives, or at least could not do so in a way that would make a difference to the arguments in the body of the paper.

Ernest Adams has long argued that the point of indicative conditionals is to express—though not to *report*—a certain feature of one's own state of mind—namely, one's assigning a high probability to the consequent, given the truth of the antecedent. At least in the interesting cases, in which the antecedent is false, indicative conditionals are, he says, without a truth-value; they are more closely akin to speech acts like 'That's disgusting!' Their primary job is the expression of a certain kind of mental state on the part of the speaker, generally with the intent to produce similar subjective states in others.[40]

Quite a few philosophers (e.g., Dorothy Edgington, Alan Gibbard, Jonathan Bennett, Richard Grandy, and Keith DeRose[41]) have developed

[40] E. W. Adams, *The Logic of Conditionals* (Dordrecht: Reidel, 1975).

[41] Dorothy Edgington, 'On Conditionals', *Mind*, 104 (1995), 235–329; Allan Gibbard, 'Two Recent Theories of Conditionals', in W. L. Harper, G. A. Pearce, and R. Stalnaker (eds.), *Ifs* (Dordrecht: Reidel, 1981); Bennett, *A Philosophical Guide*; DeRose and Richard Grandy, 'Conditional Assertions and "Biscuit" Conditionals', *Noûs*, 33 (1999), 405–20.

theories of indicatives similar to Adams's in at least this respect: assertion of an indicative conditional does not express a distinctive kind of proposition, one that is a function of the propositions expressed by antecedent and consequent; instead, it is a qualified assertion of the consequent, or the expression of a distinctive sort of mental state (e.g., 'conditional belief'). Suppose they are right, and that neither 'Eve will not sin if tempted by a toad' nor 'Eve will sin if tempted by a toad' expresses a truth. Still, there are states of mind worthy of the labels 'knowing that Eve will not sin if tempted by a toad' and 'knowing Eve will sin if tempted by a toad'. And, if indicative conditionals are supposed to be the CFs God uses in his pre-creation deliberation, these states of mind must be able to play a role in practical reasoning. On none of these theories can one truly claim to know that Eve will sin if tempted by a toad, unless one assigns a high probability to her sinning, conditionally upon her being tempted by a toad. Furthermore, one could not know this with certainty, and use the knowledge in a completely risk-free way, unless (i) the probability assigned by the knower is one, and (ii) the probability really *is* one. So, if God could truly express his state of mind, prior to creation, by saying 'Eve will freely sin if tempted by a toad', and if his knowing this conditional were supposed to give him risk-free providential control; then a theory of indicatives in this family requires that God also believe (and therefore know) that the probability of Eve's sinning, conditional upon toad-temptation, is one.

What theory of conditional probability could possibly allow for such absolutely certain and practically useful knowledge on God's part? Not a subjectivist theory (God's knowledge of the conditional probability of Eve's sinning in these circumstances must be knowledge about the world, not about his own states of mind), nor a frequentist one (the event types need never occur). I suppose the best thing one could do, were one developing a theory of Molinistic knowledge along these lines, would be to posit brute, contingent, objective 'propensities' that things can have with respect to indeterministic situations. Given the genuine indeterminacy involved, the propensities will be different in different possible worlds; and, to avoid divine determinism, it must not be up to God what they are. I believe that the arguments of this paper concerning subjunctive or counterfactual CFs could easily be transformed into arguments about contingent, objective propensities, so construed. In particular, an analogue of LIMIT would turn out to be just as implausible, and the rest of

my argument for the possibility of transworld manipulability would go through.

Neither of the remaining two families of views about indicative conditionals—theories according to which they *do* express true or false propositions constructed, at least in part, out of propositions expressed by antecedent and consequent—will provide the Molinist with a way to resist my argument.

David Lewis, Frank Jackson, and H. P. Grice argue that indicatives are simple material conditionals—but expressed in words that 'conversationally imply' much more, and thus tempt us to conclude, erroneously, that they are actually being used to assert these further things.[42] Suppose they are right. The material conditional 'If p, then q' is equivalent to 'Either not p, or q'. One might be inclined to argue as follows, against a Molinist who tries to use material conditionals instead of subjunctives: 'If Eve is tempted by a serpent, then she sins' is, on this view, equivalent to 'Either Eve is not tempted by a serpent, or she sins', a disjunction with (let us suppose) a false first disjunct and a true second disjunct. When a simple truth-functional disjunction is true, but one disjunct is false, the truth of the disjunction as a whole is dependent upon the truth of the other disjunct. Since Eve was tempted by a serpent and sinned, the truth of the disjunction, and of the material conditional, is dependent upon the fact that Eve really did sin. In general, then, true material conditionals about circumstances of free choice that actually come about are dependent upon the truth of the consequent—they are true because of what the creature in fact does. Perhaps a material conditional describing what a certain creature actually does might follow from some other truth—a truth that does not imply that the creature actually acts in the way described, such as a subjunctive conditional. But if a true material conditional does not follow from a true proposition of this sort, then its truth depends upon the truth of the antecedent—i.e., it depends upon the fact that the creature exists and acts in this way. And if the Molinist identifies the CFs about Eve, say, with material conditionals because there are no true subjunctive or counterfactual conditionals describing what Eve would freely do or would freely have done in various circumstances; then there are no other truths,

[42] Lewis, *Counterfactuals*; Frank Jackson, 'On Assertion and Indicative Conditionals', *Philosophical Review*, 88 (1979), 565–89; H. P. Grice, *Studies in the Way of Words* (Cambridge, Mass.: Harvard University Press, 1989).

independent of the fact that Eve did sin, which could imply the material conditional 'If Eve is tempted by a serpent, then she sins'. If what is true at the stages prior to God's decision to create cannot depend upon truths implying that he creates, then the material conditional about Eve and the serpent is *not* true prior to God's decision.

I recognize that this argument is not watertight. One might, for example, reasonably wonder whether the same notion of *dependence* is invoked in the two claims: true disjunctions with a false disjunct *depend upon* the truth of the remaining disjunct; and truths at the stage before God has decided to create cannot *depend upon* things implying that he creates. But the more one thinks about the idea that CFs are mere material conditionals (and that similar subjunctive conditionals and counterfactuals are not also true and available to God at the first stage), the worse it seems. There are a great many trivially true material conditionals about the free choices of possible creatures. For every possible creature x and circumstances of free choice C, if x never exists or is never in fact put in those circumstances, then material conditionals of both these forms are true: if x is in C, then x freely does A; and if x is in C, x freely refrains from doing A. It could not be the case that *every* true material conditional of this form is available prior to God's decision to create anything at all. What is true at that stage is supposed to be independent of God's decision to create any particular creatures; but, assuming that what is available at the first stage is closed under entailment, these conditionals will imply that certain possible creatures will not be created. For example, both of these are true, and they together imply that my sister does not exist: 'If my sister exists in any circumstances at all, then she will sin' and 'If my sister exists in any circumstances at all, then she will not sin'. Could some special subset of the material conditionals be available to God prior to his creative decisions? There would have to be enough CFs about me to enable God to know what will happen if he creates me, but not enough to imply that my merely possible sister will not exist. Which of the many material conditionals about my sister will God know, at the first stage?

I suppose a Molinist could imagine that a random assortment of the true CFs about my sister is somehow selected, and made available at the first stage; while, in my case, God knows the full spectrum of true material CFs about me. But there is something very strange about this idea: that, in deciding whether to create my sister, God makes use of the fact that, if my sister were offered a bribe, she would freely take it, but does *not*

make use of the fact that, if my sister were offered a bribe, she would not freely take it. After all, both are, by hypothesis, true; and true for the same reason—namely, the falsity of the antecedent. God's using just one of the two, in these circumstances, seems to me to put him in an absurd situation. It would be closely analogous to the following scenario. Suppose that, unbeknownst to me, there are neither subjunctive nor counterfactual truths about whether I would arrive at the airport on time, if I took the low road or the high road. (Perhaps quantum indeterminacy leaves it radically undetermined whether I arrive early or late, no matter which route I take.) Now, suppose I am told, by some trusted authority, that the following material conditionals are true: if I take the low road, I will arrive late; and if I take the high road, I will arrive late. On the basis of this information, I decide to call the airline, tell them that I won't make the flight, and try to arrange for a later one. Would I feel cheated to learn that it was not true that, if I had taken the low road, I would have been late; nor was it true that, if I had taken the high road, I would have been late; and that the material conditionals I was fed by this authority were true in virtue of the falsity of their antecedents? You bet I would! But if God knows 'randomly chosen' material conditionals about my (merely possible) sister, and they figure among his reasons for not creating her; and if there are no subjunctives or counterfactuals to 'back them up'; then he will know, at the next stage, that he was making decisions upon precisely this sort of basis. He was tricked!

So I set aside, as a non-starter, the hypothesis that CFs are equivalent to material conditionals. What other theories construe indicatives as true or false propositions, built, in part, out of propositions associated with the antecedent and consequent? I know of none that is at all likely to deliver indicative conditionals fit to serve as the Molinist's CFs. The most plausible of such theories assume that there is an implicit subjectivity to indicatives, a relativization to what is known by the speaker or what can be taken for granted in the context of utterance. A plausible, rough-and-ready test of the acceptability of an indicative conditional, by me at a given time, is: Try adding certainty about the antecedent to the stock of other things I then believe, modifying my other beliefs, and the probabilities I assign them, 'in the most natural, conservative manner'; and then 'see whether what results from this includes a high probability for' the consequent.[43]

[43] This is Bennett's formulation of the 'Ramsey Test'; see Bennett, *A Philosophical Guide*, 29.

One explanation for the fact that this 'Ramsey Test' seems right would be: the proposition I express by an indicative conditional on a given occasion includes an implicit relativization to the set of other things I believe then. But there are more direct arguments.

Alan Gibbard drew attention to a kind of 'stand-off' in which well-informed observers accept conditionals that clash; and these situations make the subjectivity of indicatives particularly vivid. In one famous 'Gibbardian Stand-off' (modified slightly by Bennett[44]), Pete and Lora are the only two poker players left in the game; one observer sees Pete leave the room without the distressed look that always results from Pete's calling and losing; the other observer sees Lora leave the room with more money than she had earlier. The first quite properly concludes that, if Pete called, he won the last hand; the second equally reasonably concludes that, if Pete called, he lost the last hand. Neither observer is mistaken about any 'matter of fact', or relying upon misleading evidence. Assuming that, in affirming such conditionals, the observers would express propositions, there's no reason to say that one speaks truly and the other falsely; given the aptness of the conditionals, one should conclude that both are true. That would be okay, if indicatives were merely material conditionals; but on views that treat them as something more robust, such conflicts are not tolerated.[45] The salient difference between the people in a Standoff has to do with the differences in their evidence; so if both express true propositions, the difference in evidence must somehow work its way into the truth-conditions for the conditional.

A relatively crude strategy for incorporating the speaker's evidence into the meanings of indicative conditionals would be to say that an utterance of such a conditional expresses a truth if and only if the antecedent, together with other things the speaker knows at the time, entails the consequent. Stalnaker's more sophisticated theory makes use of the kind of possible-worlds semantics that has shed considerable light on subjunctives and counterfactuals, effecting an attractive unification of the two kinds of conditionals. In Stalnaker's view, every context of thought and utterance includes a set of taken-for-granted assumptions about what the world is like, which can be represented by a set of possible worlds; and an indicative conditional is true just in case the nearest of these worlds in

[44] See Bennett, *A Philosophical Guide*, 83–93.

[45] For discussion of this point, see ibid., 84.

which the antecedent is true is also a world in which the consequent is true.[46]

In the case of an utterly lonely speaker (or thinker), the relevant set of worlds simply represents things properly-taken-for-granted by the speaker (or thinker), there being no conversation partners with whom she needs to negotiate to arrive at a common stock of reasonable assumptions. So, on a theory of indicatives like Stalnaker's, the stock of indicative conditionals God knows before deciding to create depends upon what other things he knows at that stage. If the knowledge relevant to the truth of indicatives at the first stage were allowed to include everything God knows, then at the first stage God would know things that settle what sort of world he will create. Let *p* be a proposition describing the complete future of the world. If *p* is included as part of what can be taken for granted, then God would know the proposition he could express by the words: 'If triangles have three sides, then *p* is true.' So, if there is to be a stage in God's knowledge that includes an explanation of why he created anything, the indicatives included at that stage must be evaluated using much less of what God knows. With the first stage containing only things God knows that are independent of his choice to create, and no true counterfactuals or subjunctive conditionals about freely chosen actions (since the Molinist I am imagining is trying to get by with indicatives for CFs, instead), there will not be enough in the relevant set of taken-for-granted truths at that stage to determine the truth of indicative conditionals about what free creatures will do in various circumstances.

Bibliography

Adams, E. W., *Logic for Conditionals* (Dordrecht: Reidel, 1975).

Adams, Robert M., 'Middle Knowledge and the Problem of Evil', *American Philosophical Quarterly*, 14 (1977), 109–17.

____ 'An Anti-Molinist Argument', in *Philosophical Perspectives*,v, ed. J. Tomberlin (Atascadero, Calif.: Ridgeview, 1991), 343–53.

Basinger, David, *The Case for Freewill Theism: A Philosophical Assessment* (Downers Grove, Ill.: InterVarsity Press, 1996).

[46] Robert Stalnaker, 'Indicative Conditionals', *Philosophia*, 5 (1975), 269–86.

Beilby, James K., and Paul R. Eddy (eds.), *Divine Foreknowledge: Four Views* (Downers Grove, Ill.: InterVarsity Press, 2001).

Bennett, Jonathan, *A Philosophical Guide to Conditionals* (Oxford: Clarendon Press, 2003).

Boyd, Gregory A., *God of the Possible* (Grand Rapids, Mich.: Baker Books, 2000).

Craig, William Lane, *Divine Omniscience and Human Freedom* (Leiden: E. J. Brill, 1990).

DeRose, Keith, 'The Conditionals of Deliberation', *Mind*, 00 (0000).

Flint, Thomas P., *Divine Providence: The Molinist Account* (Ithaca, NY and London: Cornell University Press, 1998).

____ 'A New Anti-Anti-Molinist Argument', *Religious Studies*, 35 (1999), 299–305.

____ and Alfred J. Freddoso, 'Maximal Power', in Freddoso (ed.), *The Existence and Nature of God* (Notre Dame, Ind.: University of Notre Dame Press, 1983), 81–113.

Freddoso, Alfred J., Introduction, *Luis de Molina: On Divine Foreknowledge* (Part IV of the *Concordia*), trans. Freddoso (Ithaca, NY and London: Cornell University Press, 1988), 1–81.

Gaskin, R., 'Conditionals of Freedom and Middle Knowledge', *Philosophical Quarterly*, 43 (1993), 412–30.

Gibbard, Allan, 'Two Recent Theories of Conditionals', in W. L. Harper, G. A. Pearce, and R. Stalnaker (eds.), *Ifs* (Dordrecht: Reidel. 1981).

Grice, H. P., *Studies in the Way of Words* (Cambridge, Mass.: Harvard University Press, 1989).

Hansen, Collin, 'Young, Restless, Reformed', *Christianity Today* (Sep. 2006), 32–8.

Hasker, William, *God, Time, and Knowledge* (Ithaca, NY and London: Cornell University Press, 1989).

____ 'A New Anti-Molinist Argument', *Religious Studies*, 35 (1999), 291–7.

Hunt, David P., 'Middle Knowledge: The "Foreknowledge Defense"', *International Journal for Philosophy of Religion*, 28 (1990), 1–24.

____ 'Divine Providence and Simple Foreknowledge', *Faith and Philosophy*, 10 (1993), 389–414.

____ 'A Reply to My Critics', *Faith and Philosophy*, 10 (1993), 428–38.

____ 'The Simple-Foreknowledge View', in Beilby and Eddy (eds.), *Divine Foreknowledge: Four Views*, 65–103.

Jackson, Frank, 'On Assertion and Indicative Conditionals', *Philosophical Review*, 88 (1979), 565–89.

Kvanvig, Jonathan, *The Possibility of an All-Knowing God* (New York: St. Martin's Press, 1986).

Lewis, David, *Counterfactuals* (Cambridge, Mass.: Harvard University Press, 1973).

——— 'Probabilities of Conditionals and Conditional Probabilities', *Philosophical Review*, 85 (1976), 297–315.

O'Connor, Timothy, 'The Impossibility of Middle Knowledge', *Philosophical Studies*, 66 (1992), 139–66.

Otte, Richard, 'A Defense of Middle Knowledge', *International Journal for the Philosophy of Religion*, 21 (1987), 161–9.

Pike, Nelson, 'Divine Omniscience and Voluntary Action', *Philosophical Review*, 74 (1965), 27–46.

Pinnock, Clark, et al., *The Openness of God* (Downers Grove, Ill.: InterVarsity Press, 1994).

Plantinga, Alvin, *God and Other Minds* (Ithaca, NY: Cornell University Press, 1967).

——— *God, Freedom, and Evil* (Grand Rapids, Mich.: Eerdmans, 1974).

——— *The Nature of Necessity* (Oxford: Clarendon Press, 1974).

——— 'Replies to My Colleagues', in J. Tomberlin and P. van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel, 1985), 313–96.

——— 'On Ockham's Way Out', *Faith and Philosophy*, 3 (1986), 235–69, repr. in John Martin Fischer (ed.), *God, Foreknowledge, and Freedom* (Stanford, Calif.: Stanford University Press), 178–215 (citations refer to Fischer).

Prior, A. N., *Time and Modality* (Oxford: Clarendon Press, 1957).

Rasmussen, Josh, 'On Creating Worlds Without Evil—Given Divine Counterfactual Knowledge', *Religious Studies*, 40 (2004), 457–70.

Reichenbach, Bruce, 'God Limits His Power', in David Basinger and Randall Basinger (eds.), *Predestination and Free Will* (Downers Grove, Ill.: InterVarsity Press, 1986), 101–24.

Sanders, John, *The God Who Risks: A Theology of Providence* (Downers Grove, Ill.: InterVarsity Press, 1998).

Stalnaker, Robert, 'A Theory of Conditionals', in 'Studies in Logical Theory', *American Philosophical Quarterly Monograph*, 2 (Oxford: Blackwell, 1968), 98–112.

——— 'Indicative Conditionals', *Philosophia*, 5 (1975), 269–86.

Swinburne, Richard, *The Coherence of Theism*, rev. edn. (Oxford: Clarendon Press, 1993).

van Inwagen, Peter, *An Essay on Free Will* (Oxford: Oxford University Press, 1986).

——— 'The Place of Chance in a World Sustained by God', in Thomas Morris (ed.), *Divine and Human Action* (Ithaca, NY and London: Cornell University Press, 1988), 211–35.

——— 'What Does an Omniscient Being Know About the Future?', in Jonathan Kvanvig (ed.), *Oxford Studies in Philosophy of Religion*, i (Oxford: Oxford University Press, 2008).

VanderLaan, David, 'Counterpossibles and Similarities', in Frank Jackson and Graham Priest (ed.), *Lewisian Themes: The Philosophy of David K. Lewis* (Oxford: Oxford University Press, 2004), ch. 10.

Wierenga, Edward, *The Nature of God* (Ithaca, NY and London: Cornell University Press, 1989).

3

The Contingency of Existence*

MICHAEL NELSON

Necessitarianism is the thesis that every proposition necessarily has the truth value it actually has. Necessitarianism is obviously false. The world is filled with contingency. I am working on this paper but I might have been swimming instead; you are reading this sentence, but you need not have been; the list goes on. Accepting contingency seems to lead to the claim that the very existence of familiar objects is contingent—a claim intuitive in its own right. This is because my existence seems to be dependent upon contingent happenings involving my parents; had those happenings not occurred, I would not have been. Furthermore, but for the actual non-occurrence of possible happenings, there would have been individuals that do not actually exist. For example, my parents could have had more children and, had they done so, those children intuitively would have been distinct from every actually existing entity. But there are powerful arguments that existence is necessary. In what follows I shall present those arguments. I distinguish six views that promise to account for the intuitions supporting the contingency of existence in light of these arguments, defending a view that derives from the work of Robert Adams.[1]

* Thanks to Karen Bennett, Chad Carmichael, Thomas Crisp, Troy Cross, Matthew Davidson, Michael Della Rocca, the late Gregory Fitch, Chris Franklin, Daniel Korman, Christopher Menzel, Samuel Newlands, Gregg Ten Elshof, Neal Tognazzini, Gabriel Uzquiano, Leslie Wolf, and especially Edward Zalta, who was generous in helping me think through and formulate the ideas of this paper and whose previous work on the topics of this paper have been sources of insight. I benefited from discussing this work with the members of the Department of Logic, History and Philosophy of Science at the University of Barcelona, the 2006 Northwest Philosophy Conference, where Korman gave comments, a work-in-progress reading group at the Claremont Colleges, and the 2007 Pacific American Philosophical Association meeting, where Bennett and Menzel gave comments.

[1] In particular, Robert Adams, 'Actualism and Thisness', *Synthese*, 49 (1981), 3–41.

1. The Contingency of Existence

We begin by formalizing our intuitions concerning the contingency of existence. There are two components to those intuitions. The first concerns the possible non-existence of actual existents and the second concerns the possible existence of actual non-existents. I propose the following.

The Possibility of Aliens (ALIEN): $\Diamond\exists x\neg\mathcal{A}\exists y(y = x)$
(It is possible that something exists that does not actually exist.)

The Existence of Possible Absentees (ABSENT): $\exists x\Diamond\neg\exists y(y = x)$
(Something exists that might not have existed.)

Assume a standard possible worlds semantics for modality. A primary concern will be with the exact nature of this model theory. So as to leave as many options open, for now we shall simply say that a model is a sequence $<\mathbf{W}, \mathbf{D}, \mathbf{w}^*>$, where $\mathbf{W}$ is a set of worlds—intuitively, the ways the universe might have been—$\mathbf{D}$ is a set of individuals, and $\mathbf{w}^* \in \mathbf{W}$—intuitively, the actual world of the model that corresponds to the way the universe is. (For simplicity I ignore the complication of accessibility relations between worlds. Furthermore, if we have non-logical predicates and individual constants in our language, then we must add a function Ψ that assigns to each n-place non-logical predicate F and world $\mathbf{w} \in \mathbf{W}$ a set of ordered n-tuple of individuals from $\mathbf{D}$ and a function Γ that assigns to each individual constant n and world $\mathbf{w}$ exactly one individual from $\mathbf{D}$.) Then ALIEN is true only in models $\mathbf{M}$ with a world whose domain contains an individual distinct from every individual in the domain of the distinguished actual world of $\mathbf{M}$. We can call such models *increasing models*, as they allow for (although do not require) worlds with domains larger than the domain of the distinguished actual world. ABSENT is true only in models $\mathbf{M}$ with a world whose domain lacks some individual in the domain of the distinguished world of $\mathbf{M}$. We can call such models *decreasing models*, as they allow for (although do not require) worlds with domains smaller than the domain of the distinguished actual world.[2] The

[2] I say allow for but do not require increasing and decreasing domains for the following reason. Consider a model where each world has a domain with the same cardinality, but domains with different objects. Then, if there are objects in the domain of some non-actual world distinct from the objects

formulae are thus only satisfiable if we have *varying domains*, where different worlds of a model have different domains, as opposed to *fixed domains*, where each world of a model has the same domain. This is intuitively exactly what we want to do justice to our intuition that what there is is contingent.

There are two assumptions supporting the adequacy of these formalizations in capturing the intuitions concerning the contingency of existence. The first is that the predicate 'x exists' is regimented as $\exists y(y = x)$ and the second is that the adverb 'actually' is regimented as the actuality operator $\mathcal{A}$, whose semantics and logic will be described below, and the predicate 'x actually exists' is regimented, with the help of that operator, as $\mathcal{A}\exists y(y = x)$ (read 'it is actually the case that there exists something identical to x'). (I will also assume that this is how to best regiment the predicate 'x is actual', as applied to ordinary individuals, although not as applied to states of affairs, possible worlds, or propositions.) I regard neither assumption as mandatory, although I think they are both true, and shall consider solutions to our problem that involve denying these assumption below. For now, however, we will simply adopt them as working assumptions.

The following provides a basic semantics for the operator $\mathcal{A}$.

$\mathcal{A}\phi$ is true with respect to (wrt) $\mathbf{w}$ in model $\mathbf{M}$ just in case ϕ is true wrt the distinguished world $\mathbf{w}^*$ of $\mathbf{M}$ in $\mathbf{M}$.[3]

in the domain of the actual world, ALIEN is true in that model, even though the domains are not increasing. And similar considerations apply to ABSENT and decreasing domains. So, ALIEN can be true even in non-increasing models and ABSENT can be true even in non-decreasing models. However, if a model is increasing, then ALIEN is true in that model and, if a model is decreasing, then ABSENT is true in that model.

[3] Typically double-indexing is thought to be necessary for adequately modeling an actuality operator (and 'now', the correlate of 'actually' in the case of tense). See, for example, John Crossley and Lloyd Humberstone, 'The Logic of "Actually"', *Reports on Mathematical Logic*, 8 (1977), 11–29; Harold Hodes, 'Axioms for actuality', *Notre Dame Journal of Formal Logic*, 31 (1984), 498–508; Hans Kamp, 'Formal Properties of "Now"', *Theoria*, 37 (1971), 227–73; David Kaplan, 'Demonstratives' (1977), in J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan* (Oxford: Oxford University Press, 1989), 481–564; David Lewis, 'General Semantics', *Synthese*, 22 (1970), 18–67; id., 'Index, Context, and Content', in S. Kanger and S. Ohman (eds.), *Philosophy and Grammar* (Dordrecht: Reidel, 1980), 79–100; Krister Segerberg, 'Two-Dimensional Modal Logic', *Journal of Philosophical Logic*, 2 (1973), 77–96; Frank Vlach, 1973, '"Now" and "Then": A Formal Study in the Logic of Tense Anaphora', 1973 Ph.D. diss., UCLA. Discussions with Edward Zalta, however, have convinced me that double-indexing is unnecessary. I cannot adequately discuss this issue here, although the issue is related to the discussion of the contingency of actuality in sect. 6.

We shall follow the logic of $\mathcal{A}$ proposed by Edward Zalta.[4] Call it the *Logic of Actuality* (henceforth LA).

LA1: $\mathcal{A}\phi \equiv \phi$

LA2: $\mathcal{A}\phi \rightarrow \Box\mathcal{A}\phi$

LA1 asserts the material equivalence of a formula and the *rigidification* of that formula, where the rigidification of ϕ is the result of adding $\mathcal{A}$ in front of ϕ. LA2 asserts that the rigification of a true formula is necessarily true. Although LA2 may seem initially counterintuitive, its validity can be seen by considering the truth definition of $\mathcal{A}$. The rigidification of a formula ϕ in any world $\mathbf{w} \in \mathbf{W}$ depends only on the truth value of ϕ in the distinguished actual world $\mathbf{w}^*$. If ϕ is contingent, then its truth value varies across worlds of $\mathbf{M}$. But the truth value of ϕ in $\mathbf{w}^*$, even if ϕ is contingent, does not vary; that remains fixed within $\mathbf{M}$. (Although the truth value of ϕ in the distinguished world may well vary across models.) But then the rigidification of ϕ has the same truth value in every world of $\mathbf{M}$, as, recall, the truth value $\mathcal{A}\phi$ in any world $\mathbf{w} \in \mathbf{W}$ is determined solely by the truth value of ϕ in the distinguished world $\mathbf{w}^*$. So, LA2 is true in every model. The controversial axiom is LA1 and there are logics of actuality that eschews LA1. (See the discussion of LA1 at the end of section 2.) For now, however, we will simply adopt LA1 as a working assumption.

With these assumptions, we can say that ALIEN and ABSENT capture the sense that existence is contingent. ALIEN is true only if there are non-actual worlds in which there are entities distinct from every actually existing object and ABSENT is true only if there exist entities in the actual world that do not exist in some non-actual worlds. Given the intuitive appeal of these principles, it is surprising and even unsettling how difficult it is to accommodate their truth in our formal modal systems and to offer an adequate metaphysics that accounts for their truth. Part of the problem stems from the mysteries surrounding non-existence and negative existentials in general. But there are also problems specific to the interaction of necessity, quantification, and actuality. To quote Arthur Prior, Q(uantified) M(odal) L(ogic), which promises to systematize our *de*

[4] Zalta, 'Logical and Analytic Truths That are Not Necessary', *Journal of Philosophy*, 85 (1988), 57–74 and id., 'Natural Numbers and Natural Cardinals as Abstract Objects: A Partial Reconstruction of Frege's *Grundgesetze* in Object Theory', *Journal of Philosophical Logic*, 28 (1999), 619–60.

re modal thinking, is 'haunted by the myth that whatever exists necessarily exists' and that necessarily everything necessarily exists.[5] Call this the *myth of necessary existence*.

After motivating the myth of necessary existence (sections 2 and 3), I present five strategies for dealing with it, the first of which involves rejecting the necessity of identity (section 4) and the other four of which involve accepting the myth as truth, denying ALIEN and ABSENT, and offering alternative explanations of the intuitions that seem to support those principles (section 5). Then (section 6), I present and defend my preferred solution to our problem, which aims to accept both the necessity of identity, ALIEN, and ABSENT. I compare my version of the view to several similar accounts from the existent literature, defending two of the characteristic features of my version: First, the denial of the characteristic S4 (i.e., $\Box\phi \rightarrow \Box\Box\phi$) and S5 (i.e., $\Diamond\phi \rightarrow \Box\Diamond\phi$) axioms and, second, the rejection of the contingency of actuality.

2. Technical Motivations for the Necessity of Existence

The myth of necessary existence has both technical and philosophical sources. There is danger in not considering these sources together, as any solution to the technical problems must mesh with a plausible solution to the philosophical, and vice versa. Furthermore, there is a danger that we fail to identify the true source of the myth if we focus too closely on just one source. In this section I present the technical problems and in the following section the philosophical.

The most direct way to motivate the myth of necessary existence is to note that (T) below is a theorem of S(tandard) Q(uantificational) L(ogic).

(T) $\exists y(y = x)$

(T) is valid because every SQL-interpretation has a non-empty domain and everything is identical to something—namely, itself. Next note that even the weakest standard propositional modal logics have as an inference rule the R(ule of) N(ecessitation), which allows us to necessitate any

[5] Arthur N. Prior, *Papers on Time and Tense* (Oxford: Clarendon Press, 1968), 48.

theorem and count the result as a theorem.[6] RN is highly plausible. RN is invalid only if there are theorems whose necessitations are not theorems. Assuming the soundness of our system, a theorem is logically true, in the sense of being true in every interpretation. But how could a logical truth fail to be necessary? How could the mere variation of contingent facts affect the truth value of a logical truth? It seems it couldn't. So, if ϕ is a theorem, then $\Box\phi$ is also a theorem. (Spoiler: my favored solution rejects the validity of RN, embracing the existence of contingent logical truths and falsehoods.)

These are the key steps to the following direct proof of the N(ecessity of) E(xistence).

1. $\exists y(y = x)$	Theorem SQL
2. $\Box\exists y(y = x)$	RN: 1
3. (NE) $\forall x\Box\exists y(y = x)$	U(niversal) G(eneralization): 2

UG from line 3 is the standard inference rule of SQL that tells us we can universally generalize any theorem and preserve theoremhood.

NE, the conclusion of the direct proof on line 3, is inconsistent with Absent. We can argue for this semantically. A formula is true in a model just in case it is true in the distinguished world of that model. Suppose that NE is true in $\mathbf{M}$. Then it is true in $\mathbf{w}^*$. So everything in $\mathbf{w}^*$ necessarily exists and hence exists in every world $\mathbf{w} \in \mathbf{W}$. But then there is nothing in $\mathbf{w}^*$ that might not have existed. Absent is true in a model just in case there is something in the domain of the actual world of the model that is not in the domain of some world of the model. So Absent is false in $\mathbf{M}$. So, any model that makes NE true makes Absent false. This is not surprising as NE and the negation of Absent are equivalent by the definitions of $\exists$ and $\Diamond$ and the elimination of double-negations.[7]

[6] More carefully, the rule of necessitation for QML is the following.

RN_{QML}: If $\vdash_{QML} \phi$, then $\vdash_{QML} \Box\phi$.

As every theorem of SQL is a theorem of QML, RN_{QML} entails that the necessitation of any theorem of SQL is a theorem of QML.

[7] Following the method suggested in the text, we start with NE ($\forall x\Box\exists y(y = x)$) and replace $\forall x$ with its equivalent $\neg\exists x\neg$, giving us $\neg\exists x\neg\Box\exists y(y = x)$. We then replace $\Box$ with its equivalent $\neg\Diamond\neg$, giving us $\neg\exists x\neg\neg\Diamond\neg\exists y(y = x)$. We now delete the embedded double negation, giving us $\neg\exists x\Diamond\neg\exists y(y = x)$, which is just the negation of Absent. Needless to say, our proof theory does not legitimate these steps.

A full and proper derivation of the negation of Absent from NE is ugly, long, and unnecessarily complex. We can simplify and help bring out the core moves by helping ourselves to the following inference rules.

NE is consistent with ALIEN.[8] We can see this intuitively as follows. NE is true in a model **M** just in case everything in the domain of $\mathbf{w}^*$ is a member of the domain of every world $\mathbf{w} \in \mathbf{W}$. ALIEN is true in a model **M** so long as the domain of some world $\mathbf{w} \in \mathbf{W}$ contains an object not in the domain of $\mathbf{w}^*$. That is, ALIEN is true in any growing model, including models in which every individual in the domain of the actual world exists in the domain of every possible world and some possible world has an individual in its domain that does not exist in the domain of the actual world. Such models make NE true as well. We can make this more precise as follows. Consider a model with two worlds, $\mathbf{w}^*$ and $\mathbf{w}'$, where the domain of $\mathbf{w}^* = \{1\}$ and the domain of $\mathbf{w}' = \{1, 2\}$. Both NE and ALIEN are true in this model. NE is true as everything in $\mathbf{w}^*$ (namely, 1) exists in every world $\mathbf{w} \in \mathbf{W}$ (i.e., $\mathbf{w}^*$ and $\mathbf{w}'$). ALIEN is also true, as there is a world (namely, $\mathbf{w}'$) where there is an object (namely, 2) that is distinct from every actual object, as $2 \neq 1$. So, the two formulae are consistent.

We can, however, easily extend the direct proof for NE with the following line

4. (NNE) $\Box\forall x\Box\exists y(y = x)$ RN: 3

to derive a formula (NNE) that is inconsistent with both ABSENT and ALIEN. NNE is inconsistent with ABSENT as NNE entails NE (by the T Axiom, $\Box\phi \rightarrow \phi$) and we have already seen that NE and ABSENT are incompatible. Intuitively, NNE is true only in models with fixed domains,

Def $\Diamond$: $\Box\phi \vdash \neg\Diamond\neg\phi$.

[The rule of] U[niversal] N[egation]: $\forall x\neg\phi \vdash \neg\exists x\phi$.

Def $\Diamond$ rests on the definition of $\Diamond$ (i.e., $\Diamond\phi =_{\text{def}} \neg\Box\neg\phi$). It is establishing UN that makes the proof ugly, but we can intuitively see UN's validity by noting, first, the definition of $\exists$ (i.e., $\exists x\phi =_{\text{def}} \neg\forall x\neg\phi$) and the validity of $\forall x\neg\neg\phi \equiv \forall x\phi$. We can then derive the negation of ABSENT from NE as follows.

1. $\forall x\Box\exists y(y = x)$ NE
2. $\Box\exists y(y = a)$ U[niversal] I[nstantiation]: 1
3. $\neg\Diamond\neg\exists y(y = a)$ Def $\Diamond$: 2
4. $\forall x\neg\Diamond\neg\exists y(y = x)$ UG: 3
5. $\neg\exists x\Diamond\neg\exists y(y = x)$ UN: 4

UI on line 2 is the equivalent inference rule of what is typically an axiom of SQL ($\forall x\phi x \rightarrow \phi a$).

[8] Thanks to Karen Bennett for correcting a claim that I made in earlier versions of this paper that NE and ALIEN are inconsistent. I also had a fallacious proof of this claim that I shall discuss in the following note.

whereas ALIEN is true only in models in which there are objects in the domain of non-actual worlds that are not in the domain of the actual world and hence in models with varying domains. We can make this more precise as follows. NNE is true in a model **M** just in case, for every world $\mathbf{w} \in \mathbf{W}$, everything in the domain of **w** exists in the domain of every world $\mathbf{w}' \in \mathbf{W}$. But then ALIEN is false in **M**, as ALIEN is true in a model only if there is a world of the model with an object distinct from every object in the domain of the distinguished actual world of that model. So, if NNE is true in a model, ALIEN is false.[9]

NE and NNE seem to run contrary to the sense that what there is is contingent. This is because these formulae are incompatible with the truth

[9] As with the incompatibility of NE and ABSENT, a full proof of the incompatibility of NNE and ALIEN is ugly. But it also raises interesting issues concerning the combination of QML and the logic of actuality.

To simplify what follows, notice that the negation of ALIEN—i.e., $\neg\Diamond\exists x\neg\mathcal{A}\exists y(y = x)$—is equivalent to $\forall x\mathcal{A}\exists y(y = x)$, by the definitions of $\exists$ and $\Diamond$ and the dropping of the inner double negation (i.e., $\neg\Diamond[\neg\neg]\exists x\neg\mathcal{A}\exists y(y = x)$). From now on, we shall treat these formulae as interchangeable. Now suppose that we have a varying domain semantics. Then, as we have seen above in the text, (1) below should be a theorem of our logic while (2) and (3) should not be theorems, as the first is valid and the last two are not.

1. $\Box\forall x\Box\exists y(y = x) \rightarrow \Box\forall x\mathcal{A}\exists y(y = x)$
   (i.e., NNE$\rightarrow$ $\neg$ALIEN)
2. $\forall x\Box\exists y(y = x) \rightarrow \Box\forall x\mathcal{A}\exists y(y = x)$
   (i.e., NE$\rightarrow$ $\neg$ALIEN)
3. $\exists y(y = x) \rightarrow \Box\forall x\mathcal{A}\exists y(y = x)$
   (i.e., (T)$\rightarrow$ $\neg$ALIEN)

So, our aim is to devise a proof theory that delivers these results.

Note that if we assume a fixed domain semantics, all three formulae are valid, ALIEN is invalid, and we have a simple derivation of the formulae, without any need to distinguish among them. The proof of the negation of ALIEN is as follows.

1.	$\exists y(y = x)$	theorem of SQL
2.	$\mathcal{A}\exists y(y = x)$	LA1: 1
3.	$\Box\mathcal{A}\exists y(y = x)$	LA2: 2
4.	$\forall x\Box\mathcal{A}\exists y(y = x)$	UG: 3
5.	$\Box\forall x\mathcal{A}\exists y(y = x)$	B(arcan) F(ormula): 4

BF is one of the mixing axioms governing the interaction of quantifiers and modal operators proposed by Ruth Barcan Marcus in her pioneering work on QML, which is the following: (BF) $\forall x\Box\phi \rightarrow \Box\forall x\phi$ ('Identity and Individuals in a Strict Functional Calculus of First Order', *Journal of Symbolic Logic*, 12 (1946), 3–23). For simplicity, I use LA1, LA2, and BF as inference rules instead of axioms. We can derive any of (1)–(3) from line 5 by relying on the following theorem of the logic of the material conditional: $\phi \rightarrow (\psi \rightarrow \phi)$. Life is much easier in the simplest QML+actuality, but that hardly makes it the right logic! (This proof is interesting in its own right. The weakest link is BF. Provided we

of ALIEN and ABSENT, which promised to capture that sense of contingency. But we derived NE and NNE from fairly weak assumptions: Namely, the resources of SQL and the resources of the weakest standard modal logic.

accept SQL and LA1, I see no other choice than to reject BF. It is also worth stressing that this proof does not rely on any application of RN, which sets it apart from all of the other proofs of NE and NNE I consider in the text. Of course, given a varying domain semantics, there are counter-examples to BF. In any case, to avoid unwanted consequences—like the validity of (3)—BF must be rejected.)

Life is harder with varying domains. Consider how we might prove (1). We require the resources of LA, defined above, and the T Axiom of standard propositional modal logics ($\Box\phi \rightarrow \phi$). (For simplicity, I shall use both LA1 and T as inference rules, where LA1 allows both the derivation of ϕ from $\mathcal{A}\phi$ and $\mathcal{A}\phi$ from ϕ.) We will simplify by helping ourselves to the following inference rule.

Def $\exists$: $\forall x\phi \vdash \neg\exists x\neg\phi$.

Finally, we shall use a conditional proof format, which allows us to infer $\phi \rightarrow \psi$ when we can derive ψ from the assumption of ϕ.

1. $\Box\forall x\Box\exists y(y = x)$	NNE [assumption]
2. $\forall x\Box\exists y(y = x)$	T: 1
3. $\Box\exists y(y = a)$	UI: 2
4. $\exists y(y = a)$	UI: 3
5. $\mathcal{A}\exists y(y = a)$	LA1: 4
6. $\forall x\mathcal{A}\exists y(y = x)$	UG: 5
7. $\neg\exists x\neg\mathcal{A}\exists y(y = x)$	Def $\exists$: 6
8. $\Box\neg\exists x\neg\mathcal{A}\exists y(y = x)$	RN: 7
9. $\Box\forall x\Box\exists y(y = x) \rightarrow \Box\neg\exists x\neg\mathcal{A}\exists y(y = x)$	CP: 1–8

While there may be more elegant ways of achieving the result, I suspect that the proponent of a varying domain semantics will have to employ a method of stripping off the $\Box$ and $\forall$ and then employing LA1 and RN.

But there is a problem. The basic strategy behind this proof threatens to allow us to derive the negation of ALIEN from both NE and (T), thus allowing us to derive (2) and (3) above, which we know to be a bad thing, as NE and ALIEN are consistent (as we proved in the text above) and ALIEN is evidently compatible with (T). Here's why. NE is line 2 of the above proof. So, start with line 2 and it seems that the rest of the proof moves along swimmingly, changing the last line with $\forall x\Box\exists y(y = x) \rightarrow \Box\neg\exists x\neg\mathcal{A}\exists y(y = x)$. Second, we can substitute (T) for line 3 of our proof. So, start our proof there and it seems that the rest of the proof moves along swimmingly, changing the last line with $\exists y(y = x) \rightarrow \Box\neg\exists x\neg\mathcal{A}\exists y(y = x)$. Furthermore, as our initial assumption here is a theorem, this would also provide us with a derivation of the negation of ALIEN directly.

So, we need to devise a system that permits the original first proof but blocks the last two. It is clear that the problem lies with the interaction between LA1 and the application of RN. Although a full account of this matter is beyond the scope of this paper, we can at least set out some of the issues concerning the complex interaction between LA1 and RN and, more generally, the logic of QML+actuality.

It is well appreciated, and I shall prove it in the text below, that LA1 and RN are incompatible with the existence of contingent truths. If LA1 is valid, we must find a restriction on RN. Our problem of distinguishing among the three proofs under consideration turns on the nature of this restriction. It might be initially tempting to supply a blanket ban on necessitating a formula in a line of a proof

We seem to face a choice: Learn to live with the results, deny one of our meager assumptions, or find some subtle mistake with the derivations that leaves intact the assumptions but still undermines the results. None of the choices are without their problems.

There is a second standard way of establishing NE. Consider the C(onverse) B(arcan) F(ormula), one of the mixing axioms for quantifiers and modal operators proposed by Ruth Barcan Marcus in her pioneering work on QML.

(CBF) $\Box\forall x\phi \rightarrow \forall x\Box\phi$[10]

if that line depends, however indirectly, on an instance of LA1. Such a restriction would block the defective proofs: as what corresponds to line 6 above in the imagined derivation of the negation of ALIEN from NE depends upon LA1, the application of RN at the line that corresponds to 6 would violate this ban. Similarly for the derivation of ALIEN for $\exists y(y = x)$. The problem, however, is that such a blanket ban also renders the acceptable proof invalid, as that proof also involves applying RN to a line that depends upon an application of LA1. Furthermore, I at least do not see any other way of deriving the negation of ALIEN from NNE, supposing a varying domain semantics. So, such a ban is too restrictive.

What we need is some way to exploit the fact that the antecedent of (1) is modal (i.e., necessary) while the antecedent of (2) and (3) are not. When we assume the antecedent of (1), manipulate it, including applying LA1, we can still validly apply the rule of necessitation to the lines that ultimately depend only on that initial necessary truth, while we can't with any non-modal (and hence potentially contingent) formulae. Such a series of steps gets us into trouble with the second and third proofs precisely because we are necessitating a line that has contingent truths as its source.

Here is one suggestion, following Zalta, *Principia Metaphysica* (unpublished MS). We could say that a line of a proof can be necessitated only if it only depends on modal formulae. We then block the two defective proofs, as the lines we apply RN to ultimately depend on non-modal formulae. But our problems aren't solved. LA1, it will be recalled, is an axiom. Hence, the proof of (1), properly filled out, will have as a line an instance of LA1. That line will be essential, in the sense that we cannot simply delete it and still reach our conclusion, but it will also, in all of the proofs under consideration, be non-modal. For just this reason, the second suggested restriction on RN comes to the same thing as the first suggested restriction: any line that ultimately (and crucially) depends on an application of LA1 will have among its dependencies a non-modal formula. So, the proof of (1) is also deemed, by this restriction, invalid.

So far our attempts to find a restriction on RN that legitimates the proof of (1) but not the sketched proofs of (2) and (3) have not met with success. But, if we conceive of LA1 as an inference rule—or really a pair of inference rules—instead of an axiom, then this second restriction serves us well. Once LA1 itself is taken out of the dependency base, we can say that line 6 of the first proof has only modal formulae in its ultimate dependency base—namely, only NNE—while the corresponding lines of the other two defective proofs do not. Thus, the application of RN in the first proof is legitimate while the applications in the other proofs are not, just as we want. This suggestion, however, would have to be worked out in detail. (Thanks to Christopher Menzel and Edward Zalta for extremely helpful discussions of this topic.)

[10] CBF requires that, for every two worlds $\mathbf{w}$, $\mathbf{w}'$ of any model $\mathbf{M}$, if $\mathbf{w}$ is accessible to $\mathbf{w}'$, then the domain of $\mathbf{w}$ is a superset of the domain of $\mathbf{w}'$. BF, Marcus's other mixing axiom, requires that, for every two worlds $\mathbf{w}$, $\mathbf{w}'$ of any model $\mathbf{M}$, if $\mathbf{w}$ is accessible to $\mathbf{w}'$, then the domain of $\mathbf{w}$ is a subset of the domain of $\mathbf{w}'$. Both principles are valid in the simplest QML with fixed domains.

CBF is provable from seemingly weak assumptions, similar to those employed in the above direct proof of NE,[11] and yet we can derive NE and NNE with the use of CBF.[12]

The standard reaction to these proofs is to fault (T)—$\exists y(y = x)$. This reaction comes in two varieties. First, one might reject SQL in favor of a free logic, in which (T) is not a theorem, empty domains are admissible, an interpretation need not assign a value to every individual constant, and the classical rule of generalization

(UI) $\forall x \varphi x \vdash \varphi(a)$

is replaced with the weaker inference rule

(UI_{FL}) $\forall x \varphi x \,\&\, E!(a) \vdash \varphi(a)$

where $E!x$ is a primitive existence predicate. Both the direct proof of NE and the derivation by way of proving CBF rely on the classical conception of quantification and fail given a free logic. (Neither the first premise of the direct proof—i.e., $\exists y(y = x)$—nor the first premise of

[11] Here is a standard proof of CBF. Saul Kripke, 'Semantical Considerations on Modal Logic', *Acta Philosophica Fennica*, 16 (1963), 83–94 is its source. See also Bernard Linsky and Edward Zalta, 'In Defense of the Simplest Quantified Modal Logic', *Philosophical Perspectives*, 8 (1994), 431–58; and Christopher Menzel, 'Actualism', *The Stanford Encyclopedia of Philosophy* (Summer 2005 edn.), ed. E. Zalta <http://plato.stanford.edu/archives/sum2005/entries/actualism/>.

1. $\forall y Fy \rightarrow Fx$	Theorem of SQL
2. $\Box(\forall y Fy \rightarrow Fx)$	RN: 1
3. $\Box(\forall y Fy \rightarrow Fx) \rightarrow (\Box\forall y Fy \rightarrow \Box Fx)$	Instance of D[istribution axiom]
4. $\Box\forall y Fy \rightarrow \Box Fx$	MP: 2,3
5. $\forall x(\Box\forall y Fy \rightarrow \Box Fx)$	UG: 4
6. $\forall x(\Box\forall y Fy \rightarrow \Box Fx) \rightarrow (\Box\forall y Fy \rightarrow \forall x\Box Fx)$	Instance axiom of SQL
7. $\Box\forall y Fy \rightarrow \forall x\Box Fx$	MP: 5,6

D from line 3 is the Distribution axiom of standard modal logics $\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$, which ensures that $\Box$ distributes across the material conditional $\rightarrow$. The key ingredients of this proof are the resources of SQL, D, and RN.

[12] Here's the proof.

1. $\forall x\exists y(y = x)$	Theorem of SQL
2. $\Box\forall x\exists y(y = x)$	RN: 1
3. $\Box\forall x\exists y(y = x) \rightarrow \forall x\Box\exists y(y = x)$	Instance of CBF
4. $\forall x\Box\exists y(y = x)$	MP: 2,3

Given CBF, this proof is impeccable. (Unlike the previous proofs, the application of RN on line 2 is uncontroversial.) We can again derive NNE by applying RN to line 4.

the derivation of CBF—i.e., $\forall \gamma F\gamma \rightarrow Fx$—are valid in free logics with empty domains.) Some take this to show that QML requires a free logic. Kit Fine, for example, proposed that an adequate QML for contingent beings should be based on a free logic.[13] Second, one might embrace the generality interpretation of theorems, which denies theoremhood to any open formula. Both the direct proof of NE and the derivation by way of proving CBF rely on counting a given formula as a theorem (or axiom), applying RN, and then applying the rule of UG to the result. Such a proof structure breaks down on the generality interpretation. Kripke famously proposed this as a way of blocking the above derivations of NE and NNE.[14]

I reject both of the standard reactions. First, they offer little insight into the philosophical motivations for the myth, to be considered in the next section. I think that we should expect a unified account of the technical and philosophical motivations for the myth.[15] It is a virtue of the accounts I shall consider in the latter sections of this paper, and in particular my preferred view, that they have a unified account of both the philosophical and technical motivations for the myth of necessary existence. Second, both accounts leave untouched the issue of whether or not we can develop a standard, classical quantified modal logic, as both responses are based on abandoning or altering SQL. Finally, I think that there is little independent support for adopting either a free logic or the generality interpretation.

Let's begin with the generality interpretation. A standard objectual semantics for quantified formulae defines the truth of such formulae in terms of the truth of open formulae under assignments of values for variables. So, semantically, standard quantification theory is based on the truth (under an assignment) of open sentences. The ban on open sentences in one's proof theory seems to me out of step with this semantics. Furthermore, as Harry Deutsch has shown, the solution also requires banishing individual constants.[16] Once individual constants are introduced and interpreted in the standard objectual manner, we can once again derive the offending results

[13] Kit Fine, 'Modal Theory for Modal Logic, Part I—The *De Re/De Dicto* Distinction', *Journal of Philosophical Logic*, 7 (1978), 125–56.

[14] Kripke, 'Semantical Considerations on Modal Logic'.

[15] Manuel García-Carpintero suggested to me that there is a deep connection between a free logic and Fregeanism. The philosophical motivations presented below in section 3 assume direct reference theory, which the Fregean will reject. There is thus the possibility of a unified response to the motivations for the myth of the necessity of existence grounded in an acceptance of free logic and Fregeanism. But I shall not explore such a view in this paper.

[16] Harry Deutsch, 'Logic for Contingent Beings', *Journal of Philosophical Research*, 19 (1994), 273–329.

without the use of open formulae as premises by simply replacing the first line of the direct proof with $\exists\gamma(\gamma = a)$ and the first line of the proof of CBF with $\forall\gamma F\gamma \rightarrow Fa$. But surely an expressively complete quantified modal logic will include a logic of terms.

Let's move to free logics. Free logics are typically motivated by the problems of negative existentials and fictional reference, on the one hand, and the sense of oddity over counting sentences like 'Something is identical to Bush' as logically true, on the other. Although a full discussion of the matter is out of the question here, I think that neither set of considerations motivates abandoning SQL in favor of a free logic. The literature on fictional reference is rich and complex, but there are several promising ways of dealing with the issue that are compatible with SQL. First, one might admit non-existent objects into one's ontology, allowing them to serve as the values of variables and be part of the domain quantification.[17] Second, one might claim that fictional objects exist, albeit as abstract objects. The intuition that they do not exist is to be accounted for as an intuition that they do not exist as concrete objects.[18] Finally, one might try to paraphrase sentences that apparently require ontological commitment to fictional characters into sentences that do not.[19] All of these accounts are consistent with SQL. Given the existence of these views of fictional reference, it is at least questionable whether fictional reference motivates a move to a free logic.

The second-mentioned motivation for free logics is found in the fact that sentences like 'Something is identical to Bush' have the status of logical truths in classical logics. This is because, assuming 'Bush' functions as an individual constant, such sentences have the form $\exists x(x = b)$ and interpreting this formula requires assigning an object from the domain of the interpretation as value of b and hence there is something—namely, that object—in the domain that is identical to it. This oddity is diminished, however, by noting that there is not a single object that serves as the value

[17] See Alexius Meinong, 'On the Theory of Objects', in R. Chisholm (ed.), *Realism and the Background of Phenomenology* (Glencoe, Ill.: Free Press, 1960), 76–117 and Terence Parsons, *Nonexistent Objects* (New Haven, Conn.: Yale University Press, 1980).

[18] See Saul Kripke, 'Reference and Existence', unpub. John Locke Lectures, 1973; Nathan Salmon, 'Nonexistence', *Noûs*, 32 (1998), 277–319; and Peter van Inwagen, 'Creatures of Fiction', *American Philosophical Quarterly*, 14 (1977), 299–308.

[19] See Kendall Walton, *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Cambridge, Mass.: Harvard University Press, 1990).

of the individual constant from interpretation to interpretation.[20] Although any admissible classical interpretation must assign *some* object from its domain, as empty individual constants are not allowed, it is not the case that George Bush himself is in the domain of every admissible interpretation of the sentence. Thus, although $\exists x(x = b)$ is logically true, it doesn't follow immediately that, as a matter of logic alone, Bush exists—there are admissible models that do not contain Bush in their domains.

The standard responses to the proofs so far surveyed do not satisfy. In my view the proper response to the direct proof of NE and NNE and the proof by way of CBF is to reject the applications of RN at lines 2 of each proof as illegitimate.[21] There are contingent logical truths and the key theorems of SQL employed in these proofs (i.e., $\exists y(y = x)$ in the direct proof and $\forall y Fy \rightarrow Fx$ in the proof of CBF) are examples. A full discussion and justification of this idea will have to wait until section 6. For now we can note that SQL-interpretations simply take as given a stock of individuals that serve as the values of free variables and individual constants and the range of quantifiers. But these individuals are contingent existents, at least under the intended interpretation. Hence, it should be no surprise that SQL generates contingent logical truths, as contingency

[20] This point is made in Deutsch.

[21] Faulting RN has its roots in the work of Arthur N. Prior (*Past, Present, and Future* (Oxford: Clarendon Press, 1967) and *Papers on Time and Tense* 'Logic for Contingent Beings'). In short, Prior distinguished a proposition's being not possibly false, which he analyzed in terms of there being no world in which it is false, from its being necessarily true, which he analyzed in terms of its being true in all worlds, denying that these two notions are interdefinable. Prior then claimed that singular propositions about contingent beings are neither true nor false in worlds in which those beings do not exist—he said that they are unstatable in some worlds. Such propositions can fail to be true in every world even though there are no worlds in which they are false. Prior used this distinction to argue against the validity of RN, claiming that it fails precisely for logical truths that are unstatable in some worlds. Although I too reject RN, I do not invoke Prior's distinction between weak and strong necessity. Rather, I distinguish two truth-like relations between a proposition and a possible world and argue that that distinction explains the failure of RN (see section 6 below for the details). Others who have faulted RN in the generation of the myth of necessary existence include Harry Deutsch, 'Contingency in Modal Logic', *Philosophical Studies*, 60 (1990), 89–102; id., 'Logic for Contingent Beings'; Fine, 'Postscript: Prior on the Construction of Possible Worlds and Instants', in Prior and Fine, *Worlds, Times and Selves* (London: Duckworth, 1977), 116–68; Fine, 'Plantinga on the Reduction of Possibilist Discourse', in J. Tomberlin and P. van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel, 1985), 145–86; Greg Fitch, 'In Defense of Aristotelian Actualism', *Philosophical Perspectives*, 10 (1996), 53–71; Christopher Menzel, 'Actualism, Ontological Commitment, and Possible Worlds Semantics', *Synthese*, 58 (1990), 355–89; and id., 'Singular Propositions and Modal Logic', *Philosophical Topics*, 21 (1993), 113–48. Kaplan ('Demonstratives') also argued that RN is invalid, although his arguments concerned the logic of demonstratives and not QML, and Kamp ('Formal Properties of "Now"') argued against the temporal equivalent of RN using the temporal indexical 'now'.

is built into the very items from which the interpretations are built. For SQL, contingent individuals are logically fundamental. Hence, they ground contingent logical truths.

The idea of faulting RN receives some support by considering a third proof of a problematic formula.

1. $\mathcal{A}\exists y(y = x) \equiv \exists y(y = x)$	Instance of LA1
2. $\forall x(\mathcal{A}\exists y(y = x) \equiv \exists y(y = x))$	UG: 1
3. $\Box\forall x(\mathcal{A}\exists y(y = x) \equiv \exists y(y = x))$	RN: 2

Line 3 says that necessarily everything is such that it actually exists iff it exists. Suppose the formula on line 3 is true in a model $\mathbf{M}$. Then, for every world $\mathbf{w} \in \mathbf{W}$, everything in $\mathbf{w}$ is such that it actually exists iff it exists. This is true just in case only actually existing individuals are in the domain of any possible world, which is inconsistent with Alien, and every actually existing individual is in the domain of every possible world, which is inconsistent with Absent.[22]

There is something disingenuous in presenting this as a serious argument for the necessity of existence, as the joint applications of LA1 and RN in general are known to cause problems: The validity of LA1 and the existence of contingent truths is inconsistent with the general validity of RN. Here's a semantic proof for this. Suppose that both LA1 and RN are valid. Then $\Box(\mathcal{A}\phi \equiv \phi)$ is valid. So, for every model $\mathbf{M}$, $\mathcal{A}\phi \equiv \phi$ is true in every world $\mathbf{w} \in \mathbf{W}$. Suppose, for reductio, that ϕ is contingent. Then ϕ is true in some worlds $\mathbf{w} \in \mathbf{W}$ and false in other worlds $\mathbf{w}' \in \mathbf{W}$. Now, ϕ is either true or false; this means that either ϕ is true in $\mathbf{w}^*$ or ϕ is false in $\mathbf{w}^*$. (I run both suppositions simultaneously, using '/'s to separate them.) Suppose ϕ is true/false in $\mathbf{w}^*$. Then, by the truth definition of $\mathcal{A}\phi$, for all $\mathbf{w} \in \mathbf{W}$, $\mathcal{A}\phi$ is true/false in $\mathbf{w}$. As we have supposed above that, for all $\mathbf{w} \in \mathbf{W}$, $\mathcal{A}\phi \equiv \phi$ is true in $\mathbf{w}$, it follows that, for all $\mathbf{w} \in \mathbf{W}$, ϕ is true/false in $\mathbf{w}$; for otherwise there would be a world $\mathbf{w} \in \mathbf{W}$ such that ϕ is false/true in $\mathbf{w}$ and $\mathcal{A}\phi$ is true/false in $\mathbf{w}$. So ϕ is not contingent after all. But ϕ was just an arbitrary formula. So, if LA1 and RN are both valid, there are no contingent truths.

[22] Appeal to the generality interpretation does not block this proof. Line 1 is an open formula. But read it in terms of its universal closure, as dictated by the generality interpretation, and we get line 2 of the proof. Unlike the standard proofs, there are no intermediate operations on the open formula prior to the application of UG.

This proves that the logic of actuality forces a choice between necessitarianism, denying the validity of LA1, or denying the validity of RN. I take the first choice to be closed, as I take our robust anti-necessitarian intuitions at face value. I think that the best option is to reject RN, recognize that some logical truths are based on contingent facts—with the logic of actuality, many instances of LA1 have this quality and, with SQL, logical truths like $\exists y(y = x)$ have this quality—and offer a restricted version of RN. While one might agree that there are contingent logical truths based on the logic of actuality, hence accepting the solution to the puzzle in the previous paragraph, one might still think that the logic of actuality is a special case and that it does not extend to other cases and in particular the case of SQL. (This is, for example, Zalta's view.) But once we have recognized some contingent logical truths, I believe we should be open to the existence of more and be open to the idea that rejecting RN provides a unified way of blocking all of the technical considerations motivating the myth of necessary existence.

There is, however, a third response one might take to the problem of the interaction of LA1 and RN and that is to conclude that LA1 is invalid. After all, RN is a highly plausible principle and there are provably sound and complete logics of actuality in which LA1 in not a theorem.[23]

Although I cannot here properly defend the claim, there is good reason to prefer accepting the validity of LA1.[24] In 'The Logic of "Actually"', Crossley and Humberstone distinguish logics of actuality that validate LA1 and those that do not. The theorems of the first are called the *general validities* and the theorems of the second the *real-world validities*. Crossley and Humberstone champion a logic of general validities. The key difference between the general validities and the real-world validities is the understanding of the notion of *truth in a model* and with it the notion of a logical truth. The logic of general validities requires that a formula is true/false in a model **M**, at a world **w** of **M**, rather than simply being true or false in a model **M**. The logic of real-world validities, on the other hand, operates with a notion of truth/falsity in a model, where (as we

[23] See, for example, Crossley and Humberstone, 'The Logic of "actually"'; Allen Hazen, 'Actuality and Quantification', *Notre Dame Journal of Formal Logic*, 31 (1990), 498–508; Hodes, 'Axioms for Actuality'; and Lloyd Humberstone, 'Two-dimensional Adventures', *Philosophical Studies*, 118 (2004), 17–65. Logics of actuality in which LA1 is valid are endorsed by Kaplan ('Demonstratives') and Zalta ('Logical and Analytic Truths That are Not Necessary').

[24] The considerations raised follow Zalta, 'Logical and Analytic Truths That are Not Necessary'.

assumed above) a formula is true in a model **M** just in case it is true in the distinguished world $\mathbf{w}^*$ of **M**. This difference allows a proponent of the logic of real-world validities to define the notion of validity (or logical truth) as truth in all models. Thus, the notion of truth in a model plays a role in explicating the notion of logical truth, as it does, for example, in Tarski's work and in Kripke's extension of Tarski's work to modal languages. For the proponent of a logic of general validities, however, formulae are not simply true or false in a model. So, the notion of a validity is not to be explicated in terms of truth in all models. Rather, a formula ϕ is said to be valid just in case, for every model **M** and every world **w** of **M**, ϕ is true in **M** at **w**. On this account there is a tight connection between logical truth and necessary truth—which is why RN is valid—as a formula is a logical truth only if it is true in every world of every model (as opposed to being true in the distinguished world of every model). But this seems to me to be the wrong account of logical truth, as it eschews the notion of truth in a model or truth in an interpretation, which seems to me basic to a philosophical understanding of logical truth and validity.

The issues between these two kinds of logics of actuality are complex. I do not pretend that considerations offered in the previous paragraph settle the matter. But I do think that they count in favor of the validity of LA1 and thus for rejecting RN. The logic of actuality has one class of contingent logical truths, forcing one restriction on RN. This opens the door to the possibility that SQL provides another class of contingent logical truths, forcing a further restriction on RN. To regiment our *de re* modal thinking in a way that respects our intuitions concerning the contingency of existence, we should reject RN. This provides us with a unified response to the proofs for the necessity of existence considered in this section: They involve too liberal a rule of necessitation. I shall return, in section 6, to provide a metaphysics that further justifies this idea.

3. Philosophical Motivations for the Necessity of Existence

In the previous section I discussed the technical motivations for the myth of necessary existence. While these motivations suffice for showing that our

problem needs to be taken seriously, it is important to see that our problem is not merely a technical problem. There are also non-formal, philosophical motivations for the myth. The basic philosophical motivations for the myth of necessary existence are grounded in the mysteries of what an individual's possible non-existence consists in. But the philosophical considerations that count against ALIEN differ from those that count against ABSENT. So we shall take each in turn.

We begin with ALIEN. What is it in virtue of which there could have been an individual that does not actually exist? There does not seem to be a satisfying answer, given six individually compelling assumptions. If ALIEN is true, then there is some bit of reality in virtue of which it is true; if true ALIEN has a truth-maker. But a robust sense of reality dictates that *what there is*, in the most inclusive of senses, consists entirely of actually existing entities. (This corresponds to the denial of Meinongianism and possibilism below.) Surely the bit of reality in virtue of which there could have been an individual that does not actually exist must be some *concrete individual* and not some property or non-concrete individual; that is, it must be because reality includes some concrete individual that might have existed but does not actually exist. (This corresponds to the denial of PH and CN below.) Putting these two points together, it would seem that the bits of reality in virtue of which ALIEN is true are actually existing concrete individuals. But ALIEN is true only if it is possible that there is something that is distinct from every actual individual. So, the actually existing concrete individual in virtue of which ALIEN is true is possibly distinct from every actual individual, including itself, as we have supposed absolutely everything there is actually exists! Nonsense. (This is grounded in the assumption of the necessity of identity.) So ALIEN is ungrounded, given our assumptions.

Let's focus our attention on the following five theses. (I will not consider the truth-maker assumption made explicit in the paragraph above.)

Meinongianism: There are non-existent entities.

Possibilism: There are non-actual entities.

P(latonic) H(aecceitism): There are individual essences that are ontologically independent of the individuals of which they are essences, in the sense that the individual essences could have existed unexemplified.

C(ontingency of) N(on-concreteness): There are concrete/non-concrete entities that might have been non-concrete/concrete.

C(ontingency of) I(dentity): There are contingently identical objects.

The argument against ALIEN assumes that these five theses are false. I shall say more about their contents and relationship to ALIEN in what follows. For now we can note the following. We assumed the falsity of Meinongianism and possibilism earlier, when we agreed to regiment 'x exists' as $\exists y(y = x)$ and 'x is actual' as $\mathcal{A}\exists y(y = x)$, with the logic ascribed to $\mathcal{A}$ by LA. This is because the following is a theorem of SQL.

$\forall x\exists y(y = x)$

Everything exists precisely because, given our regimentation of 'x exists', 'Everything exists' is just the above logical truth. And there is no room for entities that are not actual. Everything is actual precisely because, given our regimentation of 'x is actual', 'Everything is actual' (i.e., $\forall x\mathcal{A}\exists y(y = x)$) is a theorem of LA.[25] These two assumptions limit our ability to ground the truth of ALIEN. (The other three theses do not give rise to easy summation. I postpone their discussion until sections 4 and 5.)

The myth of necessary existence seems to force a choice between evils: Deny the intuitions in favor of ALIEN or embrace one of the above five counter-intuitive theses. Much better, I think, to avoid making such a choice. After exploring ways of solving our problem by embracing one of the five theses above, I shall show how we can have it all by accepting an Aristotelian conception of individuals. Grounding the truth of ALIEN does not require any of the five theses listed above.

Let's turn now to ABSENT. Timothy Williamson has presented an argument against ABSENT based on the following three claims.[26]

[25] Here is the proof.

1. $\mathcal{A}\exists y(y = x) \equiv \exists y(y = x)$	Instance of LA1
2. $\exists y(y = x) \rightarrow \mathcal{A}\exists y(y = x)$	Def $\equiv$: 1
3. $\exists y(y = x)$	Theorem SQL
4. $\mathcal{A}\exists y(y = x)$	MP: 2,3
5. $\forall x\mathcal{A}\exists y(y = x)$	UG: 4

(I use Def $\equiv$ on line 2 as a rule of inference.) If LA1 is invalid, of course, this proof fails. There is a subtle relationship between possibilism and the logic of actuality, exploration of which is beyond the scope of this paper.

[26] Timothy Williamson, 'Necessity and Existents', in A. O'Hear (ed.), *Logic, Thought, and Language* (Cambridge: Cambridge University Press), 235–7.

For all x,

(1) Necessarily, if x does not exist, then the proposition that x does not exist is true.
(2) Necessarily, if the proposition that x does not exist is true, then the proposition that x does not exist exists.
(3) Necessarily, if the proposition that x does not exist exists, then x exists.

(1), Williamson claims, rests on the more general claim that, for all propositions p, necessarily, p is true iff p. (2) rests on the more general claim that, for all propositions p, necessarily, if p is true, then p exists. (This in turn rests on the even more general claim that if something is such-and-such, then there exists something (that is such-and-such); the claim that existence precedes exemplification.) (3) is the assumption that singular propositions are ontologically dependent on the individuals they are singular with respect to and that there are no proper substitutes to serve as object–place constituents of singular negative existential propositions like the proposition that **o** does not exist. (3) is closely related to the denial of PH, which I shall discuss below in section 5. Williamson's conclusion is that, for all x, the proposition that x does not exist is necessarily false and so everything necessarily exists, which of course is the negation of ABSENT.

Williamson's argument relies on the same principles behind Alvin Plantinga's argument for the ontological independence of individual essences, which I shall discuss in sections 5 and 6.[27] (An *individual essence* of **o** is a property that, necessarily, **o** has and, necessarily, only **o** has.[28] An individual essence **E** is *ontologically independent* of the individual **o** that exemplifies **E** just in case it is possible for **E** to exist without **o** and hence without being exemplified at all).[29] The main difference between Williamson's argument and Plantinga's is that, whereas Plantinga assumed that ABSENT is obviously true and argued for independent individual essences employing the basic principles supporting (1) and (2), Williamson assumed that there are no independent individual essences and argued against ABSENT. (1) and

[27] Alvin Plantinga, 'On Existentialism', *Philosophical Studies*, 44 (1983), 1–20.
[28] Id., *The Nature of Necessity* (Oxford: Oxford University Press, 1974), 72.
[29] Adams (in his 'Actualism and Thisness') called ontologically independent individual essences *haecceities* and ontologically dependent individual essences *thisnesses*, preferring the latter over the former. I follow Adams's preference, but not his terminology.

(2) are common factors in both arguments. And both arguments rest on the view (which I shall argue in section 6 below to be mistaken) that (1) and (2) are inconsistent with the conjunction of ABSENT and the rejection of independent individual essences.

Once again, we seem forced to chose between evils: Deny one of Williamson's principles or reject ABSENT. Much better, I think, to avoid making such a choice. Again, I take it as a virtue of my preferred view that it allows us to do just that.

In this section I have presented two arguments against ALIEN and ABSENT. Although I maintain that both arguments are flawed, they are also powerful and their flaws subtle. Their very existence silences any idea that the myth of necessary existence is a purely technical problem to be solved with a purely technical fix. The myth has powerful philosophical backing. Any satisfying solution will be philosophical, and not just technical, in nature.

4. The Contingency of Identity

In the previous section I presented an argument against ALIEN and, more generally, against the idea that there could have been different objects than there actually are. The argument's soundness requires the N(ecessity of) I(dentity).

(NI) $\forall x \forall y[(x = y) \rightarrow \Box(x = y)]$

In this section I explore how rejecting NI and embracing the C(ontingency of) I(dentity) and C(ontingency of) D(iversity), below, can block this argument and, more generally, solve our problem of the myth of the necessity of existence.

(CI) $\exists x \exists y[(x = y) \,\&\, \neg\Box(x = y)]$

(CD) $\exists x \exists y[(x \neq y) \,\&\, \neg\Box(x \neq y)]$

I begin with general remarks on what I consider to be the best way to develop a contingent identity thesis. I then turn to the relationship between CI and CD, on the one hand, and NE, NNE, and the intuitions supporting ALIEN and ABSENT, on the other, ending the section with a critical discussion of the philosophical arguments against the intuitions supporting ALIEN and ABSENT presented in section 3.

CI and CD are radical forms of anti-essentialism. Anti-essentialists maintain that any distinction between necessity and contingency is ultimately grounded in conceptual relations. On the most powerful versions of the view—the linguistic doctrine of necessity—this distinction is ultimately grounded in the notion of logical truth. Necessities are grounded in logical truths and contingencies in non-logical truths. (If one accepts that there are analytic truths that are not logically true—'Every vixen is a female fox', which is not a logical truth because 'vixen' and 'female fox' are non-logical, is a plausible example—then the anti-essentialist might admit that there are necessities that are conceptually grounded even though they are not solely grounded in logical truths, but rather logical truths plus non-logical meaning relations.) The crucial claim is that modal distinctions like the distinction between an object contingently being human and necessarily being human ultimately are grounded in conceptual relations between ways of conceiving of or designating that object and the concept being human. The relationship between anti-essentialism and NI is complex. Although there are anti-essentialists who accept NI, I maintain that the motivating idea behind anti-essentialism leads to a rejection of NI. Here's why.

The anti-essentialist maintains that any distinction between necessity and contingency, including the distinction between an object's necessarily having a property and its merely contingently having that property, is grounded in conceptual relations and our ways of conceptualizing objects. The anti-essentialist can assign a nature to an object, as long as that nature is not particular to the object in question. We can put the point picturesquely by saying that an object's nature should not constrain the movement of that object through modal space in any way that is specific to that particular object; any other object should move through modal space in the same way. This is because modal distinctions are not grounded in reality independent of our ways of conceptualizing and designating it, says the anti-essentialist. Given this conception of anti-essentialism, it is all but irresistible to see identity as contingent. After all, if I am stuck being identical to me in every world (or at least in every world in which I exist) and you are stuck being identical to you in every world (or at least every world in which you exist), then our movement through modal space is severely constrained. Such natures are obviously particular to specific objects, as long as there is more than one object. So, NI, I maintain, is at odds with the basic spirit of anti-essentialism.

The best way to reject NI is inspired by Ruth Marcus's[30] argument that, contrary to Quine[31] and his followers, QML does not carry any philosophically worrying essentialist commitments.[32] The version of the view that Marcus develops validates NI.[33] But, as we shall see, the basic view can be developed so that NI comes out invalid.

[30] Marcus, 'Essentialism in Modal Logic', *Noûs*, 1 (1967), 90–6.

[31] Willard Van Orman Quine, 'Notes on Existence and Necessity', *Journal of Philosophy*, 40 (1943), 113–17; Quine, 'On the Problem of Interpreting Modal Logic', *Journal of Symbolic Logic*, 12 (1947), 43–8; and id., 'Reference and Modality', 2nd rev., in his *From a Logical Point of View* (New York: Harper and Row, 1980), 139–59.

[32] Marcus had a host of distinct arguments against Quine. For most of the others, see Marcus, 'Modalities and Intensional Languages', *Synthese*, 13 (1961), 303–22. I discuss these in my 'Anti-Essentialism and *de re* Modality' (unpublished MS).

[33] Marcus was one of the first to derive NI within second-order logic; see Marcus, 'Identity and Individuals in a Strict Functional Calculus of First Order'. Marcus offers an interesting argument that NI is consistent with anti-essentialism and that ascribing to me the property of necessarily being identical to MN is not a problematic necessary property; see id., 'Essentialism in Modal Logic', 94–5. The argument turns on the fact that, in the scope of standard non-modal logics, for any proof that involves a 'referential' premise like MN is such that he is identical to MN, there is a corresponding proof that only contains 'non-referential' premises like MN is such that he is self-identical. (I cannot adequately discuss this argument here, but I address it in my 'Anti-Essentialism and *de re* Modality'.) There is, however, an argument that, in so far as NI is valid and an object has only some of its properties necessarily, the view is not anti-essentialist in one of Quine's primary ways of understanding that thesis—namely, that any way of designating a given object is just as 'essence-revealing' as any other way of designating it. Here is the argument.

Suppose **o** in fact satisfies the condition Fx, whether uniquely or not, and suppose that $[\iota x Gx]$ (i.e., the definite description 'the G') designates **o**. (This is a weak assumption that is satisfied so long as there is some condition ϕx that **o** uniquely satisfies.) Then, employing Quine's trick of gratuitously enriching a definite description (see Quine, 'Reference and Modality', 152–3), we can construct a term that designates **o** and includes the predicate Fx—namely $[\iota x(Gx \,\&\, Fx)]$ (i.e., the description 'the G that is F'). This description also designates **o**, as we have already assumed that **o** unique satisfies Gx (by assuming that $[\iota x Gx]$ designates **o**) so it uniquely satisfies $Gx \,\&\, Fx$. Now let n be any arbitrary designator of **o**. Then $[\iota x(Gx \,\&\, Fx)] = n$ is true. But, given NI, so too is its necessitation (i.e., $\Box[(\iota x(Gx \,\&\, Fx)] = n)$). For if it weren't, we would have a counter-example to NI. But then **o** necessarily satisfies the condition Fx. (By the Russellian expansion of $[\iota x \phi x]$, the above formula is equivalent to $\Box\exists x[(Gx \,\&\, Fx) \,\&\, \forall y((Gy \,\&\, Fy) \rightarrow x = y) \,\&\, x = n]$, which entails $\Box Fn$.) So, for every condition Fx and every object **o**, **o** satisfies Fx (if and) only if **o** necessarily satisfies Fx. Thus, we get a collapse of the distinction between an object's necessarily and contingently having a property. This is bad news for Marcus's response to Quine.

It would be natural for Marcus to respond to this argument by insisting that all designators are not created equally: ordinary definite descriptions are one thing and tags, genuine proper names, or rigid designators are quite another. The necessity of a true identity statement with only tags and rigid designators, but not true identity statements with ordinary definite descriptions, is guaranteed to be true, given NI. This response, however, is inconsistent with the claim that every way of designating an object is just as essence revealing as any other way of designating that object and thus carries a commitment to Aristotelian essentialism in all of Quine's sense. But I see no other way out of the above argument. So, Marcus must either reject NI, embrace the collapse between an object's necessarily and contingently satisfying a condition, or embrace the 'Aristotelian essentialism' of claiming that only tags are essence-revealing. Her purposes of showing that Quine was wrong to think that QML carries a commitment to a philosophically problematic form of Aristotelian essentialism are much

Marcus distinguished *non-discriminating* and *discriminating necessary properties*. Suppose **o** is necessarily **F**. Then **F** is a non-discriminating necessary property of **o** just in case every other object is also necessarily **F**; otherwise, **F** is a discriminating necessary property. For example, if anything is necessarily human, then there are discriminating necessary properties, as some objects (like my kittens) are not necessarily human (as they are not human at all). On the other hand, being necessarily either exactly 5′ 7″ or not exactly 5′ 7″ is a non-discriminating necessary property, precisely because every object necessarily is either exactly 5′ 7″ or not exactly 5′ 7″. (I assume that exact heights are unproblematic properties.)

One characterization of anti-essentialism from Quine's work is the denial that there are distinguished ways of designating an object that are more 'essence revealing' than other ways of designating that object, in the sense that the set of predicates that follow analytically from one way of designating an object apply just as necessarily to that object as the predicates that follow from any other way of designating that object (see, for example, Quine 'Reference and Modality', 155). One of Marcus's insights is that this thesis is compatible with objects having non-discriminating necessary properties. What the thesis rules out is that objects have discriminating necessary properties. The most straightforward way of ensuring that the only necessary properties any object has are non-discriminating necessary properties is to accept the following R(eduction) A(xiom), so-called as it can be viewed as providing truth conditions for syntactically *de re* modal formulae (i.e., formulae in which $\Box$ governs an open sub-formula, like $\exists x \Box Fx$) in terms of syntactically *de dicto* modal formulae (i.e., formulae in which $\Box$ governs only closed sub-formulae, like $\Box \exists x Fx$).

(RA) $\exists x \Box Fx$ is true of some object **o** just in case $\Box \forall x Fx$
(**O** is necessarily **F** just in case necessarily everything is **F**.)

RA entails that the only necessary properties any object has are non-discriminating necessary properties. Thus, we have a method of distinguishing an object's necessary properties from its contingent properties that does not run afoul of Quine's primary characterization of anti-essentialism.

better served by rejecting NI. (I explore anti-essentialist accounts that embrace the collapse in Nelson, 'Anti-Essentialism and *de re* Modality').

RA entails that NI is false.[34] If RA is true, then no object—myself included—is necessarily identical with MN, as it is hardly a necessary truth that everything is identical with MN. The same holds for every other object; given RA, nothing is necessarily identical with it. Similarly, if RA is true, then no object—yourself included—is necessarily distinct from MN, as it is not a necessary truth that everything is distinct from MN.

Consider the condition $\Box(x = a)$ (i.e., 'x is necessarily identical with a'). Given RA, an object **o** satisfies this condition just in case the universal closure of the nonmodal condition $x = a$ (i.e., $\forall x(x = a)$) is necessarily true. But it is not. So, for every object **o**, whether or not **o** satisfies $x = a$, **o** does not satisfy the condition $\Box(x = a)$. Similarly for the condition $\Box(x \neq a)$. Given RA, an object **o** satisfies that condition just in case $\Box\forall x(x \neq a)$. But $\Box\forall x(x \neq a)$ is false. So, for every object **o**, whether or not **o** satisfies $x \neq a$, **o** does not satisfy $\Box(x \neq a)$. So, if RA is true, neither identity nor diversity are necessary. Hence, if RA is true, CI and CD are true.

This is not to say that anything is contingently self-identical or possibly self-diverse. Everything is necessarily self-identical, given RA, precisely because, necessarily, everything is self-identical. Distinguish NI from the N(ecessity of) S(elf-)I(dentity).

(NSI) $\forall x\Box(x = x)$

RA entails NSI. The universal closure of $x = x$—namely, $\forall x(x = x)$—is necessarily true. Interestingly, the universal closure of $x = y$—namely, $\forall x\forall y(y = x)$—is not. So, RA dictates that the first is a condition everything necessarily satisfies and the second is a condition no pair of objects necessarily satisfy, even when the pair is an identity pair and so the same object is the value of both x and y. In any event, RA dictates that NSI is true and NI is false.

We now have a better idea how to accommodate the truth of CI and CD and reject NI. The key is to accept the reduction axiom RA, which has its moorings firmly in the anti-essentialist thesis that modal attributes are not discerning across particulars. Let's apply this account, then, to the myth of necessary existence.

[34] Accepting RA provides a way of making sense of CI and CD that is independent of counterpart theory, multiple or otherwise. This fact calls into question David Lewis's claim that only counterpart theory can accommodate the contingency of identity. See David Lewis, *On the Plurality of Worlds* (Oxford: Basil Blackwell, 1986), 263.

We begin with NE (i.e., $\forall x \Box \exists y(y = x)$), which is true only if, for every object **o**, **o** satisfies the condition $\Box \exists y(y = x)$ and hence necessarily satisfies the condition $\exists y(y = x)$. RA dictates that an object necessarily satisfies a condition just in case the universal closure of that condition is necessarily true. The universal closure of our condition (namely, $\forall x \exists y(y = x)$) is necessarily true. (Proof: $\forall x \exists y(y = x)$ is a theorem of SQL; so, by RN, $\Box \forall x \exists y(y = x)$ is a theorem of QML. This is an unproblematic application of RN.) So, RA dictates that NE is true. And, as we saw in section 2, the truth of NE suffices for the falisty of ABSENT. Strictly speaking, RA is silent as regards to the validity of NNE (i.e., $\Box \forall x \Box \exists y(y = x)$). But, given the verdict on NE, it is hard to see a principled reason to reject NNE. And the truth of NNE, as we also saw in section 2, suffices for the falsity of both ABSENT and ALIEN (assuming the validity of LA1). So, let's assume our anti-essentialist accepts both NE and NNE.

How does this help, you ask, with our problem of the myth of the necessity of existence? Validating NE and NNE hardly seems a step in the right direction. But our anti-essentialist has in the offing elegant explanations of the underlying intuitions supporting ALIEN and ABSENT, even though those principles themselves are rejected.

Regarding ABSENT, we intuit, for example, that I might not have existed. We imagine how things would have been had nothing been human—surely a genuine possibility. We naturally conclude that, in such a case, I would not have existed. But such a conclusion, natural as it may be, rests on the claim that I am necessarily human, which our anti-essentialist denies. Our anti-essentialist will insist that I would have still existed in a possible world in which there are no humans, albeit not as a human. We imagine how things would have been had my parents never had children together. We naturally conclude that, in such a case, I (and my siblings) would not have existed. Again, this conclusion, natural as it may be, rests on the claim that my parentage is essential to me, which our anti-essentialist denies. I would have existed in a possible world in which my parents never had children, albeit not as their child. In so far as the intuition that I might not have existed is grounded in the belief that a certain pattern of happenings is possible and that I would not have existed had those happenings occurred, the anti-essentialist has a response. The second claim—that I would not have existed had those happenings

occurred—relies on the very essentialist theses that the anti-essentialist rejects.

There might be something lacking with this explanation. Isn't there something directly evident about the claim that it could have been that nothing is identical to me? This direct intuition has yet to be explained. I have two responses on behalf of the proponent of RA. First, I doubt that there is such a direct intuition into the truth of a principle like ABSENT. Such an intuition is based on some line of reasoning as that discussed in the previous paragraph and thus dependent on essentialist assumptions. Second, even if there were such a direct intuition, I think that it too can be explained away by the proponent of RA. I am only contingently identical to MN, even though, as we have seen, given RA, I am necessary identical to something. That is, I contingently satisfy 'x is identical with MN' and necessarily satisfy 'x is identical to something', where the latter is regimented as $\exists y(y = x)$. So, like everything else, I am such that I might not have been identical to MN. Now, consider a world in which I do not satisfy the condition 'x is identical with MN'. It is natural to think that I do not exist in such a world and that nothing in that world satisfies that condition in that world—if I don't satisfy it, then who does? But, again, that intuition rests on the essentialist thesis that, necessarily, were I to exist, I would be identical to MN and that necessarily, if that condition is satisfied, then it is satisfied only by me.

Regarding ALIEN, we intuit that it could have been that there exists something that is distinct from every actually existing object. Given RA (and the validity of LA1), this is false. But the proponent of RA can still explain the underlying intuitions.

Return to the argument against ALIEN presented in the previous section. A proponent of RA can admit that it is nonsense that an object be possibly distinct from itself. The argument against ALIEN inferred this from the fact that there is an actually existing object that is possibly distinct from every actually existing object. These are very different things. Given RA, every actually existing object is possibly distinct from every actually existing object, even though none is possibly self-diverse. What this shows is that there is no need to bloat our ontology with non-existent or non-actual individuals; there is no need to say that the intuitions supporting ALIEN are to be explained in terms of the existence of entities other than actually

concrete individuals. All we need are objects that do not have any referential necessary properties.[35]

CI shows us that we can be anti-Meinongian/possibilist/PH/CN and accept the underlying intuitions supporting ALIEN and ABSENT. But it has a steep price. Many of us, myself included, are reluctant to reject NI. Although I won't argue for NI, I think it best to seek a solution that is consistent with it. The next five solutions to the problem of the myth of necessary existence are consistent with NI.

5. Four Structurally Similar Solutions

The truth of either Meinongianism, possibilism, PH, or CN allows for metaphysically distinct yet structurally similar solutions to the myth of necessary existence consistent with NI. It is important to keep in view the structural similarities without losing sight of their important metaphysical differences. The key idea behind each of the solutions is to accept NE (i.e., $\forall x \Box \exists y(y = x)$) and NNE (i.e., $\Box \forall x \Box \exists y(y = x)$), reject ALIEN and ABSENT, and offer surrogate principles that both are consistent with NE and NNE and account for the underlying intuitions concerning the contingency of existence. In each case, this is achieved by positing more entities than one might think there are from just a casual glance around. The kinds of entity posited are, in each of the four cases, very different.

The Meinongian instantiation of this common structure is likely to be the most familiar. I begin there. Meinongianism is the thesis that

[35] Although this view is inspired by Marcus, it is worth pointing out that, despite attributions of such a view to Marcus (see, for example, Linsky and Zalta, 'In Defense of the Simplest Quantified Modal Logic'; and Karen Bennett, 'Proxy Actualism', *Philosophical Studies*, 129 (2006), 263–94), it is not the solution she presents (see Marcus, 'Dispensing with Possibilia', *Proceedings and Addresses of the American Philosophical Association*, 49 (1975), 39–51; and ead., 'Possibilia and Possible Worlds', *Grazer Philosophische Studien*, 25–6 (1986), 107–33). First, as was already noted, Marcus goes out of her way to retain NI's validity. It is true that in 'Possibilia and Possible Worlds' Marcus considers appeal to her anti-essentialist view in accommodating the possibility of her non-actual but possible brother, claiming that, as far as strict logical necessity goes, any actually existing object is possibly her brother. But she is aware that the most powerful challenge stems not from possible brothers but from the possibility of something distinct from every actually existing object. She does not appeal to anti-essentialism in responding to this challenge. Rather, she appeals to her substitutional theory of quantification, insisting that it makes sense to say that there actually are things that do not actually exist, when that initial quantifier is read substitutionally. The view ends up looking remarkably Meinongian. This is how she solves the threat of BF and ALIEN entailing possibilism. She does not, as far as I know, consider ABSENT and its apparent conflict with CBF, which she also accepts.

fundamental reality includes entities that do not exist. This entails that non-existent individuals are part of our ontology and hence available for the most unrestricted of quantifiers to range over and to serve as the values of free variables. There are many flavors of Meinongianism. We shall limit ourselves only to forms of the view consistent with the validity of $\forall x \exists y(y = x)$. Such Meinongians must introduce a primitive existence predicate $E!x$ that is distinct from the predicate $\exists y(y = x)$, as, given their view, everything is in the extension of the latter but not everything exists. Let's label this the distinction between *being* and *existence*. Given this distinction, our Meinongian can insist that our formulation of ALIEN and ABSENT fails to capture the intuition that what *exists* is contingent, as those formulations concern the contingency of *being*. Our Meinongian can then insist that the following principles better capture the intuitions concerning the contingency of *existence*.

ALIEN$_{\mathrm{M}}$: $\Diamond \exists x(E!x \,\&\, \neg \mathcal{A}E!x)$
(It is possible that [there is something that exists and does not actually exist]).

ABSENT$_{\mathrm{M}}$: $\exists x(E!x \,\&\, \Diamond \neg E!x)$
(There is something that [exists and might not have existed]).

Given the distinction between $E!x$ and $\exists y(y = x)$, ALIEN$_{\mathrm{M}}$ and ABSENT$_{\mathrm{M}}$ are consistent with NE and NNE. Our Meinongian can say, in a single breath, that, necessarily, everything that *is* necessarily *is*, but it is not necessary that everything that *exists* necessarily *exists*. Being is necessary while existence is contingent.

Meingongianism strikes me as deeply flawed—and precisely because of the posited difference between being and existence. The view is also powerful, holding the promise of accounting for the contingency of existence without exacting any radical revisions of the simplest QML—I do not take the addition of a primitive logical predicate like $E!x$ as a radical revision—and quite resilient. The standard objections to Meingonianism, in so far as those objections aim to be more than just bold assertions of the denial of Meinongianism, seem to me not to stick. The view thus deserves our attention, even if we ultimately reject it.

We can abstract the following structure from the account. We shall see that the following three solutions share the same form. Consider the following schemas.

1. $\exists x \neg \phi x$
2. $\forall x \exists y (y = x)$
3. $\Box \forall x \Box \exists y (y = x)$
4. $\Diamond \exists x (\phi x \,\&\, \neg \mathcal{A} \phi x)$
5. $\exists x (\phi x \,\&\, \Diamond \neg \phi x)$

(2) shows that each theory stays within the basic framework of SQL, where (2) is a theorem. (3) is what I have labeled NNE. Each solution is consistent with NNE and hence NE. (4) and (5) are intended to capture the intuitions supporting ALIEN and ABSENT; they are the replacement for those original principles, where the replacements are intended to be consistent with NE and NNE.[36]

Replacing all occurrences of ϕx in (1)–(5) with $E!x$ renders a consistent set of formulae, given Meinongianism. There is no need, our Meinongian says, to find fault with the derivations of NE and NNE considered above in section 2 in order to respect the intuitions that existence is contingent. Once we carefully distinguish being and existence, we can see that NE and NNE are perfectly compatible with the intuition that existence (as opposed to being) is contingent.

The CN solution is the most closely analogous to Meinongianism. Bernard Linsky and Edward Zalta defend CN as an actualistic interpretation of the simplest QML, in which both the Barcan formulas, NE, and NNE are valid.[37] The distinctive thesis of CN is that some entities are contingently concrete and others contingently non-concrete. What common intuition would count as an alien, the proponent of CN, says is an actually existing contingently non-concrete individual; what common intuition would count as an actually existing individual that might not have existed, the proponent of CN, says is an actually existing contingently concrete individual. Everything necessarily exists and anything that could exist necessarily (and so actually) exists. Our intuitions to the contrary are based on a confusion of concreteness and non-concreteness, which are genuinely contingent, for existence and non-existence, which are necessary.

[36] In each case (4) could be strengthen into (4*).

4*. $\exists x (\neg \phi x \,\&\, \Diamond \phi x)$

This is because, given (3), the domain of each world is the same and so if it is possible for a thing to exist, then that thing exists in every world. But (4) will serve our purposes well enough and has the benefit of being structurally the same as ALIEN, which it is intended to replace.

[37] Linsky and Zalta, 'In Defense of the Simplest Quantified Modal Logic'.

Replacing all occurrences of ϕx in (1)–(5) with the primitive concreteness predicate $C!x$ renders a consistent set of formulae, given CN. This demonstrates the close analogy between the Meinongian and CN solutions to the problem of the myth of necessary existence, proving that they are competing interpretations of the same set of formulae—namely, (1)–(5). But this similarity should not blind us to the important metaphysical differences between the two views. Linsky and Zalta are clear of their rejection of the distinctive Meinongian thesis.

> Just read the quantifier $\exists$ of the language of *QML* as 'there exists' or 'there is'. By actualist lights, these mean the same. Moreover, let us suppose that everything that exists is actual. This squares the object language with the thesis of actualism. Since the quantifer ranges over everything in domain **D** in the models of *QML*, everything in **D** therefore both exists and is actual.[38]

Let's assume such an anti-Meinongian version of CN. Then, unlike the Meinongian, our proponent of CN does not and cannot claim that NNE really only concerns necessary *being* rather than necessary *existence*, as that is a distinction without a difference. Rather, the proponent of CN accepts that being/existence is necessary and seeks to respect ordinary intuitions regarding possible existence and non-existence in terms of possible concreteness and non-concreteness. This allows the proponent of CN to replace Alien and Absent with the following instances of (4) and (5).

$\textsc{Alien}_{\textsc{CN}}$: $\Diamond\exists x(C!x \,\&\, \neg\mathcal{A}C!x)$
(It is possible that [something exists that is concrete and is not actually concrete]).

$\textsc{Absent}_{\textsc{CN}}$: $\exists x(C!x \,\&\, \Diamond\neg C!x)$
(There exists something that [is concrete and might have been non-concrete]).

Unlike the original Alien and Absent, $\textsc{Alien}_{\textsc{CN}}$ and $\textsc{Absent}_{\textsc{CN}}$ are consistent with NNE. And, the proponent of CN insists, the latter adequately account for the intuitions that led us to accept the former.

Let's work through two examples to better appreciate the account. Take an instance of our intuition of the possibility of aliens. We intuit that my parents could have had more children than they actually had. Suppose we

[38] Ibid., 448. Linsky and Zalta claim that 'there exists' and 'there is' *mean the same* by actualist lights. That certainly seems too strong. Even if the actualist claims that the two are coextensive, this is not because the two expressions above mean the same thing.

accept genetic essentialism and so we insist that no actually existing human person could have been those children and of course no non-human could have been their child either. (If we reject this, we are best to follow the anti-essentialist account considered above in section 4.) We conclude that, although nothing that actually exists could have been that child, it could have been that my parents had another child, in which case there would have been something distinct from everything that actually exists.[39] The proponent of CN thinks this line of reasoning faulty, as there actually exist individuals that could have been my parent's unconceived child, although they are all non-concrete. Each is only contingently non-concrete and each is such that, necessarily, had it been concrete, then it would have been my parent's child. Our intuition that such creatures do not actually exist is based on mistaking their actual non-concretness for non-existence. Search the universe up and down and we do not uncover anything that could have been my parent's unconceived child. But, claims the proponent of CN, that is because our search was only of the actually concrete individuals, which are but a small portion of the individuals that actually exist. Had our search included the contingently non-concrete individuals as well, then we would have found, among the actually existing individuals, what we were looking for.

We also intuit that I, for example, might not have existed. The proponent of CN claims that this is false. What is true in its place is that, although I am actually concrete, I might have been non-concrete. Had I been non-concrete, I would not have been something that you would find, look where you may. Once again, we confuse this true claim for the false claim that I would have not existed. I would have existed, just not among the class of concrete individuals in that world.

CN provides a coherent, elegant, and powerful solution to the problem of the myth of the necessity of existence. But I claim that we should ultimately reject that solution and precisely because of its foundational claim that concreteness and non-concreteness are contingent properties. I do not pretend to have arguments that the characteristic theses of Meinongianism and CN are false. I simply claim to intuit their falisty and bet that many will share that sense. In both cases I find direct arguments

[39] I assume, for simplicity, that there is no way to single out a particular possible child by relation only to actually existing individuals. If you think that there is a unique individual that bears a relation to, say, an actually existing sperm and ovum, then we will need to change the case.

against the views largely unconvincing. In the case of CN, there have been attempts to align the view with its possibilist relative, insisting that CN is possibilism in disguise. More recently, Karen Bennett argues that CN should be rejected as it carries a commitment to mere *actualia*—that is, entities that are actual but do not exist.[40] Both charges fail to stick: there are no entities that are non-actual and there are no entities that do not exist, according to Linsky and Zalta, but only *actually existing* individuals that are contingently concrete and contingently non-concrete.[41]

The versions of Meinongianism and CN presented above are consistent with the claim that absolutely everything is actual. They also both conceive of the domain of the most unrestricted quantifiers as being filled only with individuals and are consistent with the claim that there are only object-dependent individual essences; neither view requires a departure from standard QMLs. The next two solutions to the myth of necessary existence to be considered involve rejecting one of these assumptions: the possibilist rejects the first and the proponent of PH rejects the second. Both views, however, are consistent with the rejection of the characteristic theses of Meinongianism and CN.

We begin with possibilism. The possibilist maintains that fundamental reality includes non-actual objects. There are Meinongian and anti-Meinongian versions of possibilism. We shall focus solely on the anti-Meinongian version. A proponent of this view maintains that everything exists but some things are non-actual. David Lewis is a paradigm proponent of such a view.[42]

We face a problem in formalizing this view. Earlier we agreed to regiment the predicate 'x is actual' as $\mathcal{A}\exists y(y = x)$, where LA is the logic of $\mathcal{A}$. But $\forall x\mathcal{A}\exists y(y = x)$ is a theorem of LA[43] and so the characteristic possibilist

[40] Karen Bennett, 'Proxy Actualism'.

[41] Zalta and I discuss Bennett's critique in detail in Nelson and Zalta, 'Bennett and "Proxy Actualism"', *Philosophical Studies* (2007).

[42] Lewis, *On the Plurality of Worlds*.

[43] The proof is straightforward.

1. $\exists y(y = x) \equiv \mathcal{A}\exists y(y = x)$	Instance LA1
2. $\exists y(y = x) \rightarrow \mathcal{A}\exists y(y = x)$	Def $\equiv$: 1
3. $\exists y(y = x)$	Theorem SQL
4. $\mathcal{A}\exists y(y = x)$	MP: 2,3
5. $\forall x\mathcal{A}\exists y(y = x)$	UG: 4

Given the validity of LA1 and allowing open formulae to count as theorems, this proof is beyond reproach.

claim—that there are non-actual objects, regimented as $\exists x \neg \mathcal{A} \exists y (y = x)$—is a contradiction. The possibilist faces a choice: She can claim that 'x is actual' is not equivalent to $\mathcal{A}\exists y(y = x)$ or she can develop a logic of actuality without LA1 as an axiom. Although both options are live, it is simpler for our purposes to only consider the first, introducing a primitive logical predicate $A!x$ for 'x is actual', keeping LA as the logic of actuality, where the operator $\mathcal{A}$ is not used to explicate the predicate 'x is actual'. With this, our possibilist can claim that replacing all occurrences of ϕx in (1)–(5) with $A!x$ renders a consistent set of formulae. In particular, Alien and Absent are replaced with the following.

Alien$_P$: $\Diamond \exists x \neg A!x$
(There could have been something that is non-actual).

Absent$_P$: $\exists x (A!x \,\&\, \Diamond \neg A!x)$
(There is/exists something that [is actual and might have been non-actual]).

Alien$_P$ and Absent$_P$ are consistent with NE and NNE and are intended to account for the underlying intuitions concerning the contingency of existence. The key idea is to see fundamental reality as including individuals that are contingently actual and other entities that are contingently non-actual. What shifts from world to world is which entities are in the extension of our primitive predicate $A!x$, not what entities are in the domain of the most unrestricted of quantifiers.[44]

Again, we have an internally coherent and powerful solution to our problem that none the less strikes me as implausible, precisely because of the implausibility of taking non-actual entities to be fundamental. Many people, myself included, find it intuitive that absolutely everything there is is actual.

Let's turn to the final of our four structurally similar solutions—PH. The distinctive claim of the proponent of PH is that reality includes actually existing individual essences of every individual there could be. Intuitively, there could have been objects that do not actually exist. To accommodate, the proponent of PH claims that there are individual essences that, although

[44] Lewis would depart from our possibilist at this point. He interprets his quantifiers as restricted, worldly quantifiers. 'There are talking donkeys' is, for Lewis, false, even though, for Lewis, reality includes talking donkeys, as the quantification is implicitly restricted to this-worldly beings. So, for Lewis, the range of ordinary language quantifiers does shift from world to world. In the text I assume that we are concerned only with unrestricted quantification.

actually unexemplified, could have been exemplified. The original intuition is explained in terms of the latter possibility. Intuitively, there are objects that might not have existed. To accommodate, the proponent of PH claims that there are individual essences that, although actually exemplified, could have been unexemplified. The original intuition is explained in terms of the latter possibility. To do the work required of them, individual essences must be conceived as ontologically independent of the individuals that exemplify them. This is because every individual essence of a contingently existing being must be capable of existing unexemplified.

So far PH is in keeping with the view developed by Alvin Plantinga.[45] But there are several distinct ways of developing these ideas into a solution to the myth of necessary existence and it is not always clear which Plantinga intended. I think that there are at least three distinct ways to go.

One way to invoke independent individual essences is as follows. Whereas the domain of a possible world is standardly conceived as the set of individuals that exist in that world, the proponent of PH conceives the domain of a possible world as a class of independent essences. As such essences are necessary existents, the same class of independent essences is the domain of every world. So, our proponent of PH embraces a fixed-world semantics for QML, accepts both NE and NNE, and rejects both ALIEN and ABSENT. She then seeks to explain our intuitions concerning the contingency of existence by introducing a primitive predicate $I!x$, read 'x is exemplified', of individual essences. She claims that what varies from world to world is not the domain of the most unrestricted quantifiers—that remains fixed across worlds—but rather the extension of $I!x$. NNE is true only in models in which the same *individual essences* exist in every world and hence entails that individual essences are necessary existents. But this should not be confused with the necessary existence of *individuals*. The existence of an individual is expressed in the language of PH in terms of the exemplification of an individual essence. So, as long as different individual essences are in the extension of this primitive predicate $I!x$ from world to world, the proponent of PH has rendered a fixed-domain semantics for QML consistent with the contingency of existence of individuals. Individual essences are all necessary beings, in this ontology, while the individuals that exemplify them are not. The contingency of the latter is

[45] Plantinga, *The Nature of Necessity* and id., 'On Existentialism'.

explained in terms of the possible exemplification or non-exemplification of the former.

Replacing all occurrences of ϕx in (1)–(5) with $I!x$ renders a consistent set of formulae, given PH. The relevant instance of (4) and (5) then become the following.

ALIEN$_{PH}$: $\Diamond\exists x(I!x \,\&\, \neg\mathcal{A}I!x)$

ABSENT$_{PH}$: $\exists x(I!x \,\&\, \Diamond\neg I!x)$

ALIEN$_{PH}$ and ABSENT$_{PH}$ are consistent with NE and NNE. Furthermore, the truth of these principles, it is claimed, account for the intuitions concerning the contingency of existence.

The four views considered above share a common structure and strategy. The Meinongian claims that there is a single set of individuals that is the domain of each world, where what varies from world to world is which of those individuals exist. The proponent of CN claims that there is a single set of individuals that is the domain of each world, where what varies from world to world is which of those individuals are concrete. The possibilist claims that there is a single set of individuals that is the domain of each world, where what varies from world to world is which of those individuals are actual. Finally, the proponent of PH claims that there is a single set of individual essences that is the domain of each world, where what varies from world to world is which of those essences are exemplified. While the structure is the same, the metaphysics invoked by the views are importantly distinct. It is one thing to say, for example, that reality includes entities that do not exist, as the Meinongian does, and quite another to say that reality includes entities that are not actual. And it is one thing to say, as the proponent of CN does, that reality includes contingently concrete and contingently non-concrete individuals, and quite another to claim that reality includes unexemplified individual essences. While all of these additions to reality may offend a 'robust sense of reality', they do so in different ways and there is no reason to think that there is a single argumentative form that will do away with them all.

Before concluding this section I wish briefly to consider a problem the proponent of PH faces. Sorting it through will uncover two alternative ways of invoking independent essences to solve the myth of necessary existence that are distinct in form from the four strategies considered so far. First, the problem. According to a standard semantics, singular terms

like 'George' designate individuals and a simple sentence like 'George is a human' is true wrt **w** just in case the designation of 'George' wrt **w** falls under the extension of '*x* is human' wrt **w**. That is, each world/individual constant pair is assigned an individual, non-logical predicates are assigned sets of individuals (or sets of sequences of individuals for *n*-place predicates where $n \leq 2$) for each world, and the atomic sentence $\ulcorner n \text{ is } F \urcorner$ is true (wrt **w**) just in case the value of *n* in **w** is a member of the set assigned to $\ulcorner x \text{ is } F \urcorner$ in **w**. Furthermore, there is a deep connection between the truth of such a sentence and the truth of a corresponding quantified sentence like 'Someone is human'. $\exists x \phi x$ is true wrt **w** just in case there is some object **o** in the domain of **w** that satisfies the open formula ϕx wrt **w**. Recall that the proponent of PH conceives the domain of a world as a set of individual essences, not individuals, and hence it is individual essences that are the values of free variables and in the range of quantifiers. Individual essences, however, do not fall under predicates like '*x* is human'. It is the *individual* George, not his individual essence, that is in the extension of '*x* is human'. So there is a mismatch between the standard theory of predication and the theory of quantification presupposed by the proponent of PH, where individual essences populate domains.[46]

There are at least three ways to go in the face of this problem. The first is to abandon the claim that the domain of a possible world is a set of individual essences, returning to the traditional conception where domains are populated with individuals. Independent individual essences are then invoked, on this view, only to say what the possible non-existence of an object and the possible existence of an actually non-existent individual consist in; they play no role in the model theory for QML. An object **o**'s possible non-existence is explained, on this version of the view, in terms

[46] Here's how Linsky and Zalta express a related worry: 'A second problem is that Plantinga's modal semantics abandons our ordinary ways of thinking in nonmodal cases. Ordinarily, for "$\exists x Px$" expresses the fact that some object exemplifies property *P*. However, for Plantinga, it expresses the fact that some essence is coexemplified with *P*, and we are left without a way to express the fact that an *individual x* exemplifies a property. In fact, Plantinga's entire logic of coexemplification must be disconnected from the traditional logic of exemplification that captures our orindary ways of thinking, for if coexemplifications were "witnessed" by facts of the form *x exemplifies P* (having individual *x* as a constituent), the sentence "$\Diamond \exists P x$" would imply the existence of individual witnesses in the domains of other worlds, thus reintroducing possibilia.' 'In Defense of the Simplest Quantified Modal Logic', 442. Although I am sympathetic to Linsky and Zalta's worry, it is driven by an Aristotelian conception of the priority of individuals that the proponent of PH should—and, as I hope to show, must—reject. Without a reason to prefer the Aristotelian conception over the broadly Platonic conception supporting Plantinga's theory, this worry should not move a proponent on PH.

of the possible truth of the proposition <**E**, UNEXEMPLIFIED>, where **E** is **o**'s independent individual essence. Such a proposition's existence does not require the existence of **o** itself, as **E** can exist without **o**. Yet **E** determines **o** in every possible world that it determines anything at all. Thus propositions with **E** as a constituent can be viewed as surrogates for singular propositions with **o** itself as a constituent. Furthermore, the possible existence of something distinct from every actually existing individual is explained, on this version of the view, in terms of the possible truth of the proposition <**E**, EXEMPLIFIED>, where **E** is an actually unexemplified independent individual essence.

It might seem tempting to read Plantinga as conceiving of individual essences as playing only this limited role, not being employed directly in the semantics of QML. Plantinga seems to not accept NE and NNE[47] and, as we have seen, the most straightforward way of having individual essences play a role in the model theory validates these principles. But the evidence against such a reading of Plantinga's position far outweighs any evidence in its favor. First, Thomas Jager has developed a semantics where only individual essences populate domains,[48] which Plantinga has explicitly endorsed as capturing his intentions.[49] Second, and more profoundly, the limited role view is susceptible to the very objection Plantinga raised against Kripke's standard semantics—what Plantinga calls the *Canonical Conception*.[50] Simplifying, the Canonical Conception views a model as a set of worlds **W** with a distinguished actual world **w***, and, for each world $\mathbf{w} \in \mathbf{W}$ a domain of objects **D**, which corresponds to the individuals that exist in **w**. The individuals that constitute these domains are just ordinary physical objects. Because these entities are contingent existents, the domains of different worlds are, in some cases, distinct. But then, worries Plantinga, the union of the domains of all of the worlds in **W** has individuals that are not in the domain of the distinguished actual world **w*** and hence, he claims, do not exist. As the model theory makes free use of these

[47] There are multiple places where Plantinga asserts the truth of ALIEN; for example, *The Nature of Necessity*, 131–2 and id., 'Actualism and Possible Worlds', *Theoria*, 3 (1976), 142. And in those same passages Plantinga clearly commits himself to the truth of ABSENT, as he claims that ordinary objects, as opposed to numbers, properties, propositions, and God, are not necessary existents.

[48] Thomas Jager, 'An Actualist Semantics for Quantified Modal Logic', *Notre Dame Journal of Formal Logic*, 3 (1982), 335–49.

[49] Plantinga, 'Self-profile', in J. Tomberlin and P. van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel, 1985), 92.

[50] Plantinga, 'Actualism and Possible Worlds'.

entities, the theory carries an ontological commitment to them. Thus, claims Plantinga, the Canonical Conception entails that there are or could have been non-existent objects. Plantinga worries that this is incompatible with what he considers to be actualistic commitments.[51] In so far as these worries are sound, they arise just as much for the above described limited role view.

We should look for other ways to solve our problem of the mismatch between the standard theory of predication and the theory of quantification suggested by PH. I shall describe two, providing us, together with the above rejected view, three theories that invoke independent individual essences to solve the problem of the myth of necessary existence. Both of the following views accept a theory of quantification according to which quantifiers range over individual essences and reject the standard theory of predication according to which an atomic sentence *Fa* is true in **w** just in case the designation of *a* in **w** is a member of the set assigned to the predicate '*x* is F' in **w**. I shall first present the non-standard theory of predication common to both and then discuss their differences.

Let $\mathbf{I}_G$ be George's individual essence. We rely on a primitive relation of coexemplification between properties. Intuitively, two properties are so related when there is an individual that exemplifies both. This, of course, will not be due as an analysis of the notion, precisely because it involves direct quantification over individuals, which the present views aim to eschew. This is why I say the notion is taken as a primitive. We can then say that the atomic sentence 'George is human' is true in **w** just in case $\mathbf{I}_G$ is coexemplified with the property of humanity in **w**. Such a truth definition can be seen as the basis of an indirect theory of predication as attributions of properties to individuals is explicated in terms of coexemplification of individuals essences and ordinary properties. Individuals themselves do not make a direct appearance. Any theory where individual essences populate the domains of one's model theory must appeal to such an indirect theory of predication.

We shall say that a model is a set of worlds $\mathbf{W}$, a distinguished actual world $\mathbf{w}^* \in \mathbf{W}$, a set of individual essences $\mathbf{D}$, and a function Ψ that assigns to each *n*-place predicate *F* and world **w** a set of ordered *n*-tuple of individual essences from $\mathbf{D}$, which intuitively are the individual essences

[51] Ibid., 142.

coexemplified with **F** in **w**. (If we have individual constants, we also need an assignment function of individual constants and worlds to individual essences. I again suppress accessibility relations between worlds.) A formula $\Phi(x_1, \ldots, x_n)$ is true of a sequence of individual essences $<\mathbf{I}_1, \ldots, \mathbf{I}_n>$ in **M** and wrt **w** just in case $<\mathbf{I}_1, \ldots, \mathbf{I}_n> \in \Psi(\Phi,\mathbf{w})$. We can then define truth for quantified formulae in the standard way. For example, $\exists xFx$ is true in **M** wrt **w** just in case for some individual essence $\mathbf{I} \in \mathbf{D}$, $\mathbf{I} \in (\Psi, (Fx, \mathbf{w}))$.

A similar view has been developed by Thomas Jager.[52] There are, however, two points to make about Jager's system in relation to the view sketched above. First, Jager directly defines $\exists xFx$ as being true in a world **w** just in case there is an individual essence that is coexemplified with **F** in **w**.[53] This is neither necessary nor desirable. In the previous paragraph I employed an objectual treatment of quantification, appealing to the indirect theory of predication to define the notion of satisfaction or of an open formula being true of an individual essence. Second, there are several differences, some of them merely notational but some of them crucial, in the set up of a model. One notational difference is that Jager's models do not specify a distinguished actual world. But Jager goes on to define truth for the pair of a model and a world, presumably where that world is treated as the actual world. A more important difference, however, concerns Jager's use of the sets of essences exemplified in **w**, for each $\mathbf{w} \in \mathbf{W}$. I shall return to this difference, which is related to Plantinga's notion of an *essential domain* for a world, below.[54]

I claimed that there are two ways of invoking the indirect theory of predication in an account of the myth of necessary existence. Both views operate with a theory of quantification according to which necessarily existing independent individual essences populate domains, but some of those essences are only contingently exemplified or unexemplified. The differences between the views is that NE and NNE are valid in the first but not the second. We can trace this difference in turn to whether or not the indirect theory of predication governs the semantics of $=$.

NE and NNE come out as logical truths on such a semantics if we insist that $=$ is not subject to the indirect theory of predication. As $=$ is a piece of logical vocabulary, we define it directly in setting up our language.

[52] Jager, 'An Actualist Semantics for Quantified Modal Logic'.
[53] Ibid., 337.
[54] Plantinga, 'Actualism and Possible Worlds', 156.

One method is to say that $x = y$ is true of a sequence $<\mathbf{I}_i, \mathbf{I}_j>$ just in case that sequence is an identity pair of individual essences. On this view, the logical symbol = expresses a relation between individual essences, not individuals. Its semantics is standard. Then NE (i.e., $\forall x \Box \exists y(y = x)$) is true in every model because, for any model **M**, the domain of $\mathbf{w}^*$ consists of only individual essences all of which are in the domain of every world $\mathbf{w} \in \mathbf{W}$. NNE (i.e., $\Box \forall x \Box \exists y(y = x)$) is true in every model because, for any model **M**, the domain of every world $\mathbf{w} \in \mathbf{W}$ consists of only individual essences all of which are in the domain of every world of **W**. We then, following what was said above about PH, appeal to the contingency of the predicates $I!x$, pointing to the truth of ALIEN_{PH} and $\text{ABSENT}_{\text{PH}}$, to explain the intuitions that what there is is contingent.

The second view applies the indirect theory of predication to all predicates, including =.[55] Then we would interpret even the logical predicate = as applying, albeit indirectly, to individuals. $x = y$ is then true of the sequence $<\mathbf{I}_1, \mathbf{I}_2>$ just in case the identity relation is coexemplified with that sequence. As different individual essences are exemplified in different worlds, NE and NNE are then invalid. Consider the following model **M**. Let $\mathbf{W} = \{\mathbf{w}^*, \mathbf{w}_1\}$, the domain of both $\mathbf{w}^*$ and $\mathbf{w}_1 = \{\mathbf{I}_1, \mathbf{I}_2\}$, and $\mathbf{I}_1$ be exemplified in $\mathbf{w}^*$ but not in $\mathbf{w}_1$ and $\mathbf{I}_2$ be exemplified in $\mathbf{w}_2$ but not in $\mathbf{w}^*$. Then NE is false in **M**, as $\mathbf{I}_1$ does not satisfy $\Box \exists y(y = x)$ as $\mathbf{I}_1$ is not exemplified in $\mathbf{w}_2$. NNE is likewise false in **M**.

There are two points to be made about this second theory. First, we cannot simply add to its language a primitive logical predicate $I!x$ corresponding to 'x is exemplified', where that predicate is intended to express a property of individual essences. This is because such a predicate will be subject to the indirect theory of predication and hence will end up ascribing a property to individuals. We can add such a predicate to our object-language, of course, but only by either introducing a new level of entities used to ascribe properties to individual essences or by explicitly excluding that predicate from the indirect theory of predication governing other predicates and invoking a direct theory of predication for it. Second, on this view the domains of the worlds of a model are the same set of individual essences, as in fixed-domain semantics for QML. But the

[55] While I am not sure the view I articulate below is quite what he had in mind, I owe its inspiration to discussions with Thomas Crisp.

semantic consequences of a standard fixed-domain semantics—namely, the validity of NE and NNE—are avoided precisely by the adoption of the non-standard theory of predication, even for logical predicates like =. Because of this benefit, plus the demonstrated need for a proponent of independent individual essences to appeal to the indirect theory of predication, I take this to be reason to prefer the second theory to the first. And, of course, this view does not share a structural similarity to the four views that were the primary focus of this section.[56]

6. Aristotelianism

In the previous two sections I considered five solutions to the problem of the myth of necessary existence. While I have not offered arguments against them, none seems attractive, precisely because of the metaphysics they presuppose. In this section I shall show how we can respect our intuitions concerning the contingency of existence by accepting both ALIEN and ABSENT and rejecting NE and NNE, without positing special entities that one would not recognize from a casual glance around the universe. The key is to embrace Adams's and Fine's distinction between how matters stand in a world from how matters stand at a world. This distinction provides a philosophical basis for the rejection of RN, the unrestricted Rule of Necessitation, which I identified in section 2 as the problematic source of the derivations of NE and NNE, and will provide responses to the philosophical arguments against the contingency of existence. We thus get a unified solution to the various strands of the problem we have been discussing.

[56] In 'Actualism and Possible Worlds', 156 Plantinga wonders what role domains play in a QML for contingent beings. He worries that sets of individuals do not serve the purpose of modeling formulae of QML without entailing that there are or at least could have been individuals that do not exist. Plantinga introduces the notion of the *essential domain* of a world **w**, which is the set of individual essences exemplified in **w**. He claims that these play the role of domains, which is to provide entities that are the range of quantifiers and the values of free variables and individual constants. Neither view presented above in the text invoke essential domains—they only invoke sets of independent individual essences, whether exemplified or not. It seems to me a mistake to think of the domain of a world as an essential domain. On Plantinga's view, even at worlds in which an individual essence is unexemplified, that individual essence is implicated in certain formulae true in that world. For example, consider a me-less world. It is true in that world that I do not exist. For Plantinga, this is because, in that world, my individual essence is not exemplified. So my individual essence had better be around in that world, otherwise we violate the principles Plantinga worried Kripke's standard varying domain semantics violates. Essential domains do not serve the formal needs of domains. They only tell us what individuals exist at a world.

To help explicate the in/at distinction, we begin by describing what I shall call an Aristotelian conception of individual essence, according to which an individual is more fundamental than its individual essence. While the historical Aristotle seems to have had little truck with individual essences, preferring to conceive all essences as purely general properties like humanity and statuehood,[57] the title for this view is none the less fitting, for at least two related reasons. First, Aristotle maintained that individuals are primary substances, where primary substance is the category of things whose existence is required for the existence of other categories of things, like properties. Second, Aristotle maintained that properties in general are dependent upon the individuals that exemplify them. Both of these theses are in the same spirit as the Aristotelian conception of individual essences.[58] The Aristotelian includes individuals among the fundamental grounds of necessity and contingency.

The following considerations seem to me to motivate, although not conclusively support, Aristotelianism. The only ordinary properties we can count on playing the role of individual essences are referential identity properties like the property being identical with MN. The argument for this is based on the possibility that there is nothing but two qualitatively indiscernible objects, **a** and **b**.[59] Because **a** is distinct from **b**, **a**'s individual essence is distinct from **b**'s individual essence. These individual essences are thus not purely qualitative, as any quality that **a** has **b** has as well and vice versa. There seem to be two choices. The first is to introduce a distinct class of primitive properties that are stipulated to be unshareable and yet not constructed out of other, more familiar materials. While this is not the option Plantinga actually takes, I believe that it is the option he is in the end forced to take. According to this view, **a**'s individual essence is an unanalyzable property that necessarily **a** exemplifies and necessarily only **a** exemplifies. That property is primitively distinct from **b**'s individual essence. The problem with this view is that it makes individual essences

[57] An individual substance like Socrates is individuated in terms of its general essences—which, in the case of Socrates is humanity and is shared with other individual substances like Glaucon—and the matter of which it is composed. It is their differing matter which distinguish Socrates from other substances, like Glaucon, with the same general essences. It is not implausible to maintain that Socrates's and Glaucon's differing matter are individuated in virtue of their differing spatio-temporal locations.

[58] See Fitch, 'In Defense of Aristotelian Actualism', 57 for similar reasons for calling his defended view 'Aristotelian actualism'. Thanks to Troy Cross for suggesting to me that the issue between Plantinga and Adams is an instance of the issue between Platonism and Aristotelianism.

[59] Max Black, 'The Identity of Indiscernibles', *Mind*, 61 (1952), 153–64.

mysterious. The second is to identify **a**'s individual essence with a more familiar non-qualitative property: **a**'s referential identity property. Because **a** is diverse from **b**, the property being identical to **a** is exemplified by **a** and not **b**. Furthermore, assuming the necessity of identity and diversity, this is an individual essence of **a**. Ditto for the property being identical to **b** for **b**. It is hard to see any other plausible candidate for individual essences of **a** and **b**.[60] Such properties are constructions from individuals. On the most plausible view, the property is constructed by lamba-abstraction on the first argument place in the proposition <IDENTITY,(**a**, **a**)>. (Alternatively, we could conceive of the property as being constructed by saturating the second argument place in the identity relation with **a**.) The property thus is dependent on **a** and so would not have existed had **a** not existed. It seems, then, that, in general, independent individual essences are not going to be any ordinary property that we already accept. The only ordinary properties that have any hope of playing the role of individual essence are referential identity properties, which are ontologically dependent on their exemplifiers. So, in so far as individual essences are not mysterious properties, the Aristotelian thesis must be true of them.

Recall that Plantinga employed individual essences to explain an individual's possible non-existence, saying that **o**'s possible non-existence consists in the proposition that **E** is not exemplified being possibly true (where **E** is an individual essence of **o**). This explanation is not available to the Aristotelian, as **E** would not have existed if **o** had not existed. And this is where the truth-in/truth-at distinction comes in.

Consider two worlds, $\mathbf{w}_{@}$ and $\mathbf{w}_{1}$. Let the first be the actual world and contain **o** as a member of its domain. Let the second not contain **o** as a member of its domain. The singular proposition <EXISTS, **o**> does not exist in $\mathbf{w}_{1}$ and hence is neither true nor false in $\mathbf{w}_{1}$. This is because a constituent of that proposition—namely, **o**—does not exist in $\mathbf{w}_{1}$ and propositions are

[60] **a** and **b**'s differing locations might provide an exception. One might say that **a**'s individual essence is the property of being in location **l** at **t** in **w**. (It is crucial that we conceive of the location itself being a constituent of this property and not some purely qualitative characterization of that location, lest we get a property that **b** satisfies as well. We also need to rigidify, directly including **w**, lest objects be necessarily nailed to their locations.) As only **a** exemplifies in **w** the property of being in location **l**, only **a** exemplifies the rigidification of that property in any world. But this property, it might be claimed, is independent of the existence of **a** itself, in the sense that it could have existed even if **a** did not. It would take us too far afield to discuss this in detail.

dependent up their constituents.[61] Appealing to surrogates for this singular proposition that contain individual essences will not help, for the reason given in the previous paragraph. But in $\mathbf{w}_@$, that proposition both exists and is true. From the the perspective of $\mathbf{w}_@$ and using the resources available in that world, we can say that $\mathbf{w}_1$ is a world at which <EXISTS, o> is false. That is to say, that proposition is false at $\mathbf{w}_1$, which is just what it takes for that proposition to be possibly false and hence only contingently true.

The truth-in/truth-at distinction corresponds to two ways of evaluating a proposition relative to a possible world. Let's conceive of a possible world as a maximal and consistent set of propositions,[62] constructed only from entities that actually exist. In conceiving a world we might constrain ourselves to the propositions that would exist were it actual. In our case neither the proposition <EXISTS, o> nor its negation would exist and

[61] Plantinga claims that the notion of constituency this claim is based on is obscure ('On Existentialism', 7–9). He writes: 'If [*sic*] feel as if I have a grasp on this notion of constituency when I'm told that, say, wisdom but not beauty is a constituent of the proposition Socrates is wise; but when it is added that Socrates himself is a constituent of that proposition, I begin to lose my sense of what's being talked about. If an abstract object like a proposition has constituents, wouldn't they themselves have to be abstract? But secondly: if we're prepared to suppose something as initially outré as that persons can be constituents of propositions, why insist that a proposition is ontologically dependent upon its constituents?' (ibid., 8–9) Plantinga understands constituency for sets and agrees that sets are ontologically dependent on their members. This is an important ingredient in his argument that an actualist must reject the Canonical Conception's identification of properties with sets ('Actualism and Possible Worlds', 146). We do not need to assume that propositions are sets to work off this understanding; we need only assume that propositions are set-like, built from more basic entities by a proposition-building function much as sets are built from more basic entities by a set-building function. We can then undermine Plantinga's worry by considering the following analogous claims about sets: 'I feel as if I have a grasp on this notion of constituency when I'm told that, say, 7 but not 8 is a constituent of the set {7, Socrates}; but when it is added that Socrates himself is a constituent of that set, I begin to lose my sense of what's being talked about. If an abstract object like a set has constituents, wouldn't they themselves have to be abstract? But secondly: if we're prepared to suppose something as initially outré as that persons can be constituents of set, why insist that a set is ontologically dependent upon its constituents?' Sets are clearly ontologically dependent on their members, are clearly abstract objects, yet some sets clearly have non-abstract constituents. Given the analogy between sets and propositions, we should say the same of propositions. Plantinga may have other reasons for objecting to the idea that concrete particulars are constituents of abstract propositions, but the oft-cited passage from above does not seem to provide a compelling reason.

[62] Both Adams, 'Actualism and Thisness' and Plantinga, *The Nature of Necessity* identify possible worlds with sets of propositions. There are a number of objections to this identification. The most troubling are based on worries that paradox lurks in this conception (see, for example, Patrick Grim, 'There is No Set of All Truths', *Analysis*, 46 (1986), 186–91 and Michael Jubien, 'Problems with Possible Worlds', in D. Austin (ed.), *Philosophical Analysis: A Defense by Example* (Dordrecht: Kluwer Academic Publishers, 1988), 299–322.) Nothing I say in the text requires this identification; I adopt it merely for its simplicity. Fitch, 'In Defense of Aristotelian Actualism', 55, contains an alternative conception of possible worlds consistent with the overall view I am developing. Christopher Menzel has also developed an alternative conception.

so neither is a member of the relevant set of propositions. The set is still complete in a genuine sense, as it details a complete description of how matters would stand were it actual.[63] As a proposition is true in a world just in case it is a member of the set of propositions that would be true were it actual, neither proposition about **o**'s existence is true in $\mathbf{w}_1$. On the other hand, in conceiving a world we might employ all the resources available us. Here we don't worry about how things would be were the world actual but how the world is, from the perspective of the actual world. In our case, <NEG<EXISTS, **o**>> is a member of the set of propositions that constitute how matters stand at $\mathbf{w}_1$ and hence that proposition is true at $\mathbf{w}_1$.

This distinction undermines P[lantinga's] M[aster] A[rgument] against the Aristotelian conception of individual essence and in favor independent individual essences. (Let **G** be the proposition [*It is not the case that George exists*].)

(P1) George might not have existed (i.e., **G** might have been true).

(P2) For all propositions **p**, if **p** were true, then *p* would have existed.

(P3) Necessarily, if **G** were true, then George would not have existed.

(C) It is possibly true that (**G** is true and George does not exist).[64]

(C) entails that George's possible non-existence consists in the possible truth of a proposition that would have existed even if George did not. We should all assume that any singular proposition with George himself as constituent would not have existed had George not existed. The most plausible view is that, given the truth of (C), **G** is a proposition with George's individual essence as constituent, where individual essences are a primitive property that are stipulated to be necessarily non-communicable. Hence, if (C) is true, Aristotelianism is false, in so far as individual essences are non-qualitative. That's PMA.

[63] This isn't quite right, for reasons connected to the truth of ALIEN. But right now we are focused on ABSENT.

[64] The argument is a simplification of the argument in Plantinga's 'On Existentialism'. Plantinga's version has an intermediate premise between (P1) and (P2), moving from **G**'s being possible to its being possibly true. The only reason I can see that Plantinga takes this more scenic route is to open a space for Arthur Prior's view, according to which singular existential propositions are possible but not possibly true. Plantinga wrongly aligns Adams and Fine with Prior on this score, claiming that the former would, like the Priorian, deny this step. As I show below in the text, this is a misunderstanding of the kind of view adopted by Adams and Fine. The truth-in/truth-at distinction is not equivalent to Prior's possible/possibly true distinction and the Adams response to PMA is very different from the Priorian response Plantinga considers.

PMA turns on an equivocation between truth-in and truth-at. (P2) is true understood in terms of truth-in. For any proposition **p** and world **w**, if **p** is true in **w**, then **p** would have existed were **w** actual. But (P1) is false when understood in terms of truth-in. There is no world **w** such that **G** is true in **w**. For **G** is true in no world whatsoever; it is false in all Georged-worlds and does not exist, and hence is neither true nor false, in any George-less world. (P1) is true understood in terms of truth-at. For indeed George might not have existed, in the sense that **G** is true at some worlds. But (P2) is false understood in terms of truth-at; a proposition need not exist in a world to be true at that world. There is no univocal reading of the premises, however, according to which they are all true. So, PMA fails to motivate the thesis that individual essences are ontologically independent of the individuals that exemplify them.

Let's turn to Williamson's argument against ABSENT presented in Section 3. Although Williamson is right that his argument does not turn on an equivocation as Plantinga's argument does,[65] it does run afoul of the truth-in/truth-at distinction nonetheless. The general principle supporting Williamson's second premise is, in effect, (P2) above. Recall that that premise is only true, if true at all, when read in terms of truth-in, as it is certainly false when read in terms of truth-at. So, for the argument to be sound, the premises and conclusion must be understood in terms of truth-in. The other principles supporting Williamson's premises are true when read in terms of truth-in. Williamson takes this to bode well for his argument. After all, his argument is sound when all of the premises are understood in terms of truth-in. But then the conclusion must also be understood in terms of truth-in. So understood, however, the conclusion is compatible with the contingency of existence and the truth of ABSENT, contrary to Williamson's intentions. Recall from our discussion of PMA that we agreed that there is no world in which a negative singular existential is true. The contingency of existence and truth of ABSENT are grounded not in a negative singular existential proposition being true *in* some world but rather in its being true *at* some worlds. Necessity and contingency concern what is true at non-actual possible worlds. So, to challenge the truth of ABSENT, Williamson's conclusion

[65] See Williamson, 'Necessity and Existents', 238–9.

must be understood in terms of truth-at, in which case his argument is unsound, as it either turns on an equivocation (if all of the premises are understood in terms of truth-in) or has a false second premise (if the premises are all understood in terms of truth-at). So, the truth-in/truth-at distinction can be invoked to undermine Williamson's argument against ABSENT.

So far we have focused on ABSENT. I have argued that Aristotelianism and the in/at distinction provide a plausible account of its truth and a satisfying response to the philosophical arguments against it. Let's turn now to ALIEN. Given the conception of possible worlds we have inherited from Plantinga and Adams, *all* possible worlds are constructs from actually existing propositions. To account for the truth of ABSENT, the Aristotelian leaves some actually existing propositions (and their negations) out, so to speak, in characterizing how matters stand in that world or how things would be were that world actual. So one might think that to account for the truth of ALIEN the Aristotelian should simply add some non-actually existing propositions into the set. The problem is that we cannot add singular propositions involving non-actually existing objects, as, given actualism, there are no such propositions anywhere to be found. Consider an instance of the problem. We want $\Diamond\exists x\neg\mathcal{A}\exists y(y = x)$ to be true. So there is some world $\mathbf{w}$ accessible from $\mathbf{w}^*$ such that $\exists x\neg\mathcal{A}\exists y(y = x)$ is true at $\mathbf{w}$. If we think of the standard truth definition for quantificational formulae, we might be tempted to say that this is the case just in case there is some object $\mathbf{o}$ that satisfies the formula $\neg\mathcal{A}\exists y(y = x)$ at $\mathbf{w}$ and hence there is a singular proposition of the form $\neg\mathcal{A}\exists y(y = \mathbf{o})$ that is a member of the set of propositions specifying how matters stand in $\mathbf{w}$. But that's the rub. Given our assumptions, there is no such object, as absolutely everything there is is identical to some actually existing object, and so there is no such singular proposition.

Reality neither includes such an individual nor such a singular proposition. But it might have. What we should say is that $\exists x\neg\mathcal{A}\exists y(y = x)$ is true at $\mathbf{w}$ even though there is no instance of that formula that is true at $\mathbf{w}$ and no object $\mathbf{o}$ that satisfies the condition $\neg\mathcal{A}\exists y(y = x)$ at $\mathbf{w}$. There would have been such an instance were $\mathbf{w}$ actual. From the perspective of $\mathbf{w}$, there is an instance of that generalization that is true in $\mathbf{w}$. This means that, with our impoverished (from the perspective of $\mathbf{w}$) resources, we cannot fully represent how matters stand in $\mathbf{w}$. This should be a welcome

consequence for those who, like the Aristotelian, insist that the individuals there happen to be ground the space of possibilities.[66]

Before concluding our discussion of Aristotelianism, I shall discuss three outstanding issues. Two of them arise in connection to the above discussion of the truth of ABSENT and ALIEN. They will serve to set my position apart from other similar views in the extant literature. The first concerns the standing of the characteristic S4 and S5 axioms and the second the nature of the contingency of actuality. The third concerns the invalidity of the rule of necessitation and the needed restriction, bringing us back to the technical motivations for the myth of the necessity of existence from section 2.

While the issues are complex and I do not pretend to offer conclusive support for these claims, I think it plausible to maintain that Aristotelianism leads to the invalidity of the characteristic S4 and S5 axioms.[67]

S4: $\Box\phi \rightarrow \Box\Box\phi$

S5: $\Diamond\phi \rightarrow \Box\Diamond\phi$

As Kripke in 'Semantical Considerations on Modal Logic' demonstrated, the validity of these principles depends on the nature of the accessibility relation between the worlds of a model. If $\mathbf{w}$ is accessible from $\mathbf{w}'$, then the happenings at $\mathbf{w}$ (from the perspective of $\mathbf{w}'$) represents genuine possibilities for $\mathbf{w}'$. It seems obvious that the accessibility relation is reflexive, in so far as logical and metaphysical necessity and possibility are at stake. Any world is accessible from itself, as $\phi \rightarrow \Diamond\phi$ seems clearly valid. Given this, S4 is valid just in case the accessibility relation is transitive and S5 is valid just in case the accessibility relation is an equivalence relation and hence every world of a model is accessible from any other.

Let's consider specific examples. $\Box p \rightarrow \Box\Box p$ is an instance of S4. Let $\mathbf{M}_4$ be a model where $\mathbf{W} = \{\mathbf{w}^*, \mathbf{w}_1, \mathbf{w}_2\}$. Let p be true in $\mathbf{w}^*$ and $\mathbf{w}_1$, but false in $\mathbf{w}_2$. Finally, let $\mathbf{w}_1$ be accessible from $\mathbf{w}^*$, $\mathbf{w}_2$ be accessible from $\mathbf{w}_1$, and each world be accessible from itself; hence, $\mathbf{w}_2$ is not accessible from $\mathbf{w}^*$. Once we allow models where the accessibility relation between worlds is not an equivalence relation, we need to make explicit something in the truth definition of $\Box$ that we have left implicit. $\Box\phi$ is true in $\mathbf{M}$ just in case ϕ is true

[66] Adams, 'Actualism and Thisness' and Fitch, 'In Defense of Aristotelian Actualism' contain similar accounts of the truth of ALIEN.

[67] Adams, 'Actualism and Thisness' contains an argument that these formulae are invalid. My discussion follows his.

in all worlds $\mathbf{w} \in \mathbf{W}$ that are accessible from $\mathbf{w}^*$. $\Box\phi$ is true in $\mathbf{M}$ and with respect to a world $\mathbf{w} \in \mathbf{W}$ just in case ϕ is true in all worlds $\mathbf{w}' \in \mathbf{W}$ accessible from $\mathbf{w}$. Then we can say that $\Box p$ is true in $\mathbf{M}_4$, as p is true wrt every world accessible from $\mathbf{w}^*$—namely, $\mathbf{w}^*$ and $\mathbf{w}_1$. But $\Box\Box p$ is false in $\mathbf{M}_4$, as there is a world accessible from $\mathbf{w}^*$—namely, $\mathbf{w}_1$—where $\Box p$ is false, as there is a world accessible from it—namely, $\mathbf{w}_2$—where p is false. (We get this result because the accessibility relation is not transitive in $\mathbf{M}_4$.) So, $\mathbf{M}_4$ is a counter-example to S4. I shall argue that the considerations of the truth of ALIEN presented above together with Aristotelianism lead to such a situation.

Now let's consider a counter-example to S5. Let $\mathbf{M}_5$ be a model where $\mathbf{W} = \{\mathbf{w}^*, \mathbf{w}_1\}$. Let p be true in $\mathbf{w}^*$ and false wrt $\mathbf{w}_1$. Finally, let $\mathbf{w}_1$ be accessible from $\mathbf{w}^*$ and from itself and let $\mathbf{w}^*$ be accessible only from itself and hence not accessible from $\mathbf{w}_1$. $\Diamond\phi$ is true in $\mathbf{M}$ just in case there is some world $\mathbf{w} \in \mathbf{W}$ accessible from $\mathbf{w}^*$ such that ϕ is true wrt $\mathbf{w}$. So, $\Diamond p$ is true in $\mathbf{M}_5$ as $\mathbf{w}^*$ is accessible from itself and p is true wrt $\mathbf{w}^*$. But $\Box\Diamond p$ is not true in $\mathbf{M}_5$ as $\mathbf{w}_1$ is accessible from $\mathbf{w}^*$ and there is no world accessible from $\mathbf{w}_1$ where p is true. (We get this because the accessibility is not symmetric in $\mathbf{M}_5$.) So, $\mathbf{M}_5$ is a counter-example to S5. I shall argue that the considerations of the truth of ALIEN presented above together with Aristotelianism lead to such a situation.[68]

Whether or not the Aristotelian thesis leads to counter-examples to S4 and S5 turns in large part on 'which singular *modal* propositions, if any, should be counted as true at worlds in which individuals they are about would not exist'.[69] Adams's contention is that a proposition like [*I am possibly identical to something*] ascribes a property to me and hence should be false at worlds where I do not exist. The motivating intuition is that the only singular propositions about an individual **o** that are true at a world where **o** does not exist are those that are determined by **o**'s non-existence and non-modal propositions true in that world. To think that any others are true is to think that individuals would have positive characteristics even had they not existed, which is strange. This leads to Adams's (C6). (See p. 29; my characterization differs from Adams's.)

[68] Nathan Salmon, 'Modal Paradox: Parts and Counterparts, Points and Counterpoints', *Midwest Studies in Philosophy*, 11 (1986), 75–120 and id., 'The Logic of What Might Have Been', *Philosophical Review*, 98 (1989), 3–34 contain a different set of arguments, not based on Aristotelianism, against the validity of S4 and S5. I find Salmon's arguments very compelling and to offer independent reason for rejecting S4 and S5.

[69] Adams, 'Actualism and Thisness', 28.

For any singular proposition of the form [*It is possible that* **p**] or [*It is necessary that* **p**] about **o** and any world **w** such that **o** does not exist at **w**, [*It is not possible that* **p**] and [*It is not necessary that* **p**] are true at **w**.

This principle fits naturally with Aristotelianism and leads to the invalidity of S4 and S5.

Let's begin with the latter claim. Consider a world **w** where I do not exist that is accessible from the actual world. (We are guaranteed such a world from the fact that it is true that I might not have existed and such possibilities are a matter of how matters stand at non-actual worlds.) The proposition [*It is necessary that (if MN exists, then MN = MN)*] is false at **w**, given the above principle. So, there is a world accessible from the actual world at which it is false that, necessarily, if MN exists, then MN = MN. But it is necessary that if MN exists then MN = MN. At every world accessible from the actual world, the proposition [*if MN exists, then MN = MN*] is true. The non-modal proposition is vacuously true at **w**. (It is important to see that the above principle governs only modal propositions.) So, we have a counter-example to S4: It is necessary that, if MN exists then MN = MN, but not necessarily necessary that, if MN exists then MN = MN.

The same holds for S5. Consider a world **w** where I do not exist that is accessible from the actual world. The proposition [*It is possible that MN exists*] is false at **w**, given the above principle. Now, as I do exist, it is possible that I exist. So, we have a counter-example to S5: It is possible that MN exists, but it is not necessarily possible that MN exists.[70]

Let's turn now to our first claim above: that Aristotelianism leads naturally to this account of what singular modal propositions about **o** are true at worlds where **o** does not exist. Adams claims that the above principle is metaphysically satisfying 'from an actualist point of view, because there *are* no possibilities or necessities *de re* about nonactual individuals' and that accepting this involves opting 'for a modal logic that reflects the idea that what modal facts there are (or would be) depends on what individuals there are (or would be)' (p. 29). The satisfaction derives from the fact that the Aristotelian maintains that the individuals there happen to be, together

[70] I have not here provided an account of what propositions are true at worlds, as I agree with Adams's account (see 'Actualism and Thisness', 23–30). What I aim to do here is defend the most controversial component of that account—namely, the principles governing modal singular propositions, which is what delivers the above results concerning S4 and S5.

with their substantive natures, provide the fundamental grounding for all *de re* possibilities and necessities. And had there been other individuals than there happen to be, those individuals would have provided different grounds for different *de re* possibilities and necessities. There are no *de re* possibilities concerning individuals that do not actually exist. (Although, as the truth of ALIEN shows, there are general possibilities concerning such individuals.) And, had a given actual existent not existed, then there would not have been *de re* possibilities and necessities concerning that individual. The space of possibilities is dependent upon the individuals there happen to be.

Let's consider an alternative account of the singular modal propositions true at worlds. One such principle is the following.

(C6′) No matter what the form of **p**, any proposition of the form [*It is possible that* **p**], [*It is necessary that* **p**], [*It is not possible that* **p**], or [*It is not necessary that* **p**] is true at any world **w** just in case it is *true* (i.e., true in the actual world **w***).[71]

(C6′) eliminates the above, or any, counter-examples to S4 and S5. Thus, the issue concerning whether or not Aristotelianism leads to the invalidity of S4 and S5 turns on whether we accept the original principle for the truth of singular modal propositions at worlds or a principle like (C6′).

(C6′) is contrary to the spirit of Aristotelianism, as it violates the requirement that what is true about an individual **o** at a world where **o** does not exist is determined solely by **o**'s non-existence at that world and propositions true in that world. Meeting this requirement is necessary to ensure that individuals are not ascribed properties at worlds at which they do not exist; which seems required by the fact that individuals *are* no way when they do not exist. This is what allows us to respect the intuitions behind Plantinga's thesis of serious actualism—the view that exemplification requires existence—while still claiming that objects do not exist at some worlds.

Jason Turner has recently defended (C6′), which corresponds to Turner's (C4).[72] Turner argues that the above considerations offered against (C6′) are based on a confusion between 'the *de dicto* [[*It is possible that* **o** *has the property* **F**]] and the *de re* [[**o** *possibly has the property* **F**]]. The former says something about the *proposition*, whereas the latter makes a claim about the

[71] Adams' Actualism and Thisness, 31.
[72] Jason Turner, 'Strong and Weak Possibility', *Philosophical Studies*, 125 (2005), 205.

object' (p. 204).[73] But there is no such confusion. The modal propositions in question are *singular* propositions. The distinction between ascribing possibility to such a proposition and ascribing possible exemplification of the property involved to the individual that proposition is about collapses in such cases. There is a genuine, metaphysical difference between saying that the proposition [*The inventor of bifocals is dead*] is possible and saying that the inventor of bifocals possibly has the property of being dead. But that is only because the proposition is not singular with respect to Benjamin Franklin. Benjamin Franklin only gets into the act directly in the second, *de dicto* case. But Benjamin Franklin himself is directly involved in the ascription of possibility to the singular proposition [*It is possible that Benjamin Franklin is dead*]. (There is, of course, a *syntactic* distinction between the sentences 'It is possible that *a* is *F*' and '*A* is such that it is possible that it is *F*' that can be called the *de dicto/de re* distinction. But where *a* is a directly referential singular term and so the embedded sentence '*A* is *F*' expresses a singular proposition, there is no corresponding *metaphysical* (or semantical, for that matter) distinction concerning whether or not the object itself is being ascribed a property. And it is a metaphysical distinction that concerns us here.) So, ascribing possible truth to a proposition singular with respect to me at a world where I do not exist is equivalent to ascribing a possible property to me at that world. As the latter is deemed problematic by both sides of the controversy, we should all agree that the former is problematic as well. Our problems with (C6′) remain. The Aristotelian should accept the original principle governing the truth of modal propositions at worlds and with it the invalidity of S4 and S5.

I began this discussion with the disclaimer that the relationship between Aristotelianism and the validity or invalidity of S4 and S5 is complex and that I do not pretend to have conclusive arguments for their invalidity. But I do believe that the above considerations suggest that the metaphysical picture concerning the fundamentality of individuals and their grounding possibilities leads to counter-examples to S4 and S5 and thus to the idea that possibilities are relative.

[73] Christopher Menzel makes a similar point. In his 'Singular Propositions and Modal Logic' there is an elegant and detailed account of the truth of ALIEN and ABSENT in the spirit of the view I present in which S4 and S5 are valid. The key move is to find actually existing set-theoretic stand-ins for aliens (Greg Ray, 'An Ontology-free Modal Semantics', *Journal of Philosophical Logic*, 25 (1996), 333–61 contains a nominalist version of Menzel's account). I hope to engage with the details of Menzel's view in future work.

Let's turn to the second issue mentioned above: The contingency of actuality. What I have to say about this issue is even more tentative than what I said about S4 and S5. We begin with the notion of truth-in. I suggested above that the propositions true in a world are those that characterize how the world would have been were it actual. This is how Plantinga characterizes truth-in a world: 'To say that p is true in a world W is to say that if W had been actual, p would have been true.'[74] Plantinga went on to use this notion to define necessity and contingency: $\Box\phi$ is true just in case ϕ is true in every world; $\Diamond\phi$ is true just in case ϕ is true in some worlds. For an Aristotelian, this is the wrong account of necessity and contingency, as those notions should be explicated in terms truth-*at* a world and truth-at does not concern how matters would stand were that world actual. But the Aristotelian still employs the notion of truth-in and, as we have seen, it is crucial to his explanation of the truth of ALIEN and ABSENT. And this is what leads to our problem: What does it mean to say of a non-actual world that it could have been actual?

Both Plantinga and Adams seem to agree that the contingency of actuality is critical to avoiding necessitarianism. Adams is explicit. He writes:

> But the denial [of the claim that non-actual possible worlds are possibly actual] entails that there is no such thing as contingent actuality. We would have to conclude that the actual world, in all its infinite detail, is the only possible world that could have been actual. And we would be left to wonder in what sense the other possible worlds are possible, since they could not have been actual.[75]

The idea seems to be that possibility consists precisely in certain situations being possibly actual. If a situation, whether maximal or not, is not possibly actual, then it is not a genuine possibility at all.[76]

While there is something clearly right about this contention, I find it problematic as stated. The source of the problem is in treating the

[74] Plantinga, *The Nature of Necessity*, 46.

[75] Robert Adams, 'Theories of Actuality', *Noûs*, 8 (1974), 222. Karen Bennett (in 'Two Axes of Actualism', *Philosophical Review*, 114 (2005), 316) makes this objection to the view she dubs **@-ism**, according to which 'the thesis of actualism, whatever it is, is true given that this world @ is the actual world' (p. 311).

[76] Harry Deutsch (in 'Contingency in Modal Logic' and 'Logic for Contingent Beings') employs this idea of the contingency of actuality, along with a two-dimensional semantics inspired by Kaplan's logic of demonstratives, in his logic for contingent beings. I hope to discuss the details of this view in future work. I think, however, that the view is mistaken for its assumptions regarding the contingency of actuality and employment of a two-dimensional framework.

contingency of actuality on a par with the familiar contingency of, say, my wearing a blue shirt instead of a brown one. To explicate the familiar sort of contingency, we say that I have one property in the actual world and a contrary property in a non-actual possible world. Similarly, the picture seems to be, there is a property of actuality that the actual world has, but other worlds might have had. The contingency of actuality, however, is not to be explicated in the same terms as familiar examples of contingency. To think otherwise seems to me out of keeping with Adams's idea that all possibilities are ultimately grounded in what is actual. If non-actual possible worlds were basic, primitive entities, as they are, for example, for the possibilist, as opposed to logical constructs 'out of the furniture of the actual world',[77] then this picture of actuality seems to have a place. Similarly, if non-actual possible worlds concern which independent individual essences are coexemplified with which properties, then, again, this picture has some grip. This is because there is some sense, on both views, that the space of possibilities is set independently of the individuals there happen to be. But, as we have seen, this is not what the Aristotelian maintains.

For the Aristotelian, actuality (as applied to situations, states of affairs, propositions, or worlds) should not be conceived as a property alongside familiar properties like shape, size, and location, which vary from world to world.[78] What then are we to make of the sensible claim that non-actual possible worlds are 'possibly actual' if not in terms of the shifting of the feature of actuality from world to world? I propose that we understand it in terms of a variation across models.[79] Models are composed of both

[77] Adams, 'Theories of Actuality', 224.

[78] Two-dimensionalism provides one example of this mistake. Though there are many versions of the view, and many different projects the view has been implicated in—from the explanation of information, belief, and other propositional attitudes, to an account of the contingent *a priori* and necessary *a posteriori*, to an account of the connection between conceivability and possibility—I am here only interested in an account of metaphysical possibility. The core two-dimensionalist idea is that we have a space of worlds and two ways of 'considering' non-actual worlds: As counterfactual and as actual. The two dimensions, however, are considered within a single model. I argue against this idea in the text.

[79] Bennett writes: '[T]here *is* a sense in which other worlds *aren't* possibly actual—namely, no other world could have been actual when 'actual' is meant rigidly. However, that is not the right sense; that is not the split between possibility and possible actuality at issue . . . Rather, what is in danger of being undermined is the very intuitive link between possibility and possible actuality in the *shifty* sense' ('Two Axes of Actualism', 316). This distinction is part and parcel of the two-dimensionalist framework mentioned in the previous note. I maintain that there is only one sense of 'actual' and that's the 'rigid sense'. I attempt to do justice to the so-called shifty sense of 'actual' with the talk in the text

a set of worlds and a distinguished world from that set. The intended model has as its distinguished world the set of propositions all of which are true. But we can, once we have the sets of worlds constructed from actually existing materials, consider centering a model on another world and use the same machinery. It is this kind of variation—a shift across models—that the 'contingency of actuality' consists in. Now, if the world that we re-center on is one that verifies ALIEN, then we cannot really fully specify how matters stand in that model, for the reasons discussed earlier.

Variation across possible worlds is used to explicate necessity and contingency, evaluate counterfactuals, etc. I claim that that framework is ill suited to explicate the contingency of actuality. It is wrong to think of, within a single model, there being two ways of 'considering' a non-actual possible world: as actual and as counterfactual. There is one way and that is as counterfactual. The work being done by the idea of 'considering a non-actual world as actual' is best done by considering a new model from the intended one, employed for the jobs mentioned above of explicating necessity and contingency, evaluating counterfactuals, etc., which is centered on that non-actual world. I suggest that this distinction is critical to a proper understanding of the property of actuality and its fundamentality. Once we conceive the 'contingency of actuality' in the second, preferred way, we will not be tempted to think of actuality as *contingent* in anything like the same sense that, say, my position in space-time is contingent.[80]

of moving to different models with the same set of worlds but a different distinguished world. But this is not a true 'shiftiness' in the ordinary sense and is quite distinct from shifts across contexts or shifts across worlds.

[80] Bennett claims that the denier of the contingency of actuality must reject what has become the standard actualist solution to the problem of iterated modalities. The problem of iterated modalities is this. While nothing that actually exists could have been my brother, intuitively, I could have had a brother. Furthermore, my possible brother is such that he could have been an only child. This is taken to be a problem because our actualist maintains that reality does not include any individual that grounds this second possibility, which none the less appears to be a genuine possibility. The standard response is to claim that, while the second possibility is not a genuine possibility, it could have been a possibility had the first possibility been actual. Something like this solution to the problem of iterated possibilities is presented in Adams, 'Actualism and Thisness', 9 and Fitch, 'In Defense of Aristotelian Actualism', 64. Bennett argues ('Two Axes of Actualism', 317–18) that this response is not available to those that deny that non-actual worlds could have been actual. I have two responses. First, I doubt that the success of the response requires the contingency of actuality. What it requires is that there are possibilities not accessible from the actual world. It could have been that I had a brother and, had I had a brother, then it would have been possible that he was an only child. There is no need to say of this first non-actual

Let us turn to our third and last issue: the Rule of Necessitation. Recall the three proofs of problematic formulae discussed in section 2. I claimed that the best response is to reject RN. Aristotelianism and the distinction between how matters stand in a world and how they stand at a world provides a framework for explaining the failure of RN.[81] Some ordinary logical truths are dependent on contingent truths.[82] $\exists\gamma(\gamma = a)$ is a logical truth, given SQL. But it is not a necessary truth because it is dependent on the existence of **a**, which is contingent. $\exists\gamma(\gamma = a)$ is a logical truth because it is true in every interpretation. Moving from formulae to propositions, we can say that the existence of the singular proposition $[\exists\gamma(\gamma = \mathbf{a})]$ suffices for its truth. But it is contingent because there are worlds at which it is not true; such worlds, of course, would not be suitable models to interpret $\exists\gamma(\gamma = a)$, but the restraints on models should not be carried over to constraints on worlds. Models interpret the non-logical vocabulary of our language; worlds represent ways the universe might have been. These are different roles, with different constraints. (This point was also behind my explanation of the 'contingency of actuality'.)

If we reject RN, we still need a $\Box$-introduction rule, lest our logic be hopelessly incomplete. We do not want to necessitate logical truths that are dependent on contingent facts. Given the genuine contingency of existence, truths like those expressed by $\exists\gamma(\gamma = a)$ are examples. These formulae are not, however, valid in free logics, precisely because empty domains and empty individual constants are admissible. Now, whereas the proponent of a free logic typically purports to capture the class of truths of logic, we reject that idea, claiming that SQL does just fine on that score. We invoke free logics to provide us with a criteria for when a formula can validly be necessitated. So, we propose to resort to free logics not to determine the class of logical truths, but rather to determine, in our proof theory, which of the class of logical truths are also necessary and

possibility that it is possibly actual. Second, even if we insist that the contingency of actuality must get in the game somewhere, I have provided an explanation of the 'contingency' of actuality in terms of shifts across models that is consistent with the denial that actuality shifts across worlds. So, I maintain that the standard response to the problem of iterated modalities is consistent with the denial of the contingency of actuality.

[81] I again follow Adams, 'Actualism and Thisness'.

[82] This is familiar from the logic of demonstratives and, if we assume LA1, the logic of actuality (Kaplan, 'Demonstratives'). Kaplan's explanation of this fact, however, is that logical truth is a feature of character and necessity a feature of content. The explanation given below is different, and I think more satisfying, and applies to cases in which the formulae in question have constant characters, where Kaplan's explanation does not get traction.

hence can be validly necessitated. We can then take our restricted Rule of Necessitation to be the following:

RN*: If ϕ is a theorem of a (sound and complete) free logic, then $\Box\phi$ is a theorem.[83]

RN* provides the foundation of a QML that does justice to our intuitions concerning the contingency of existence. The theory of quantification at its heart is classical; the metaphysics presupposed by the semantics eschews entities at odds with a 'robust sense of reality', such as non-existents, mere *possibilia*, contingently concrete and contingently non-concrete individuals, and independent individual essences. While the view has its costs—in particular, the logic is much more complicated and less powerful than the simplest QML, for example—I think that, on balance, this view deserves best in show.

Conclusion

We have motivated the problem of the myth of necessary existence. The myth has robust support, both technical and philosophical. We surveyed six solutions to this problem and suggested that the best is one grounded in an Aristotelian conception of the fundamentality of individuals. The most distinctive features of this account is that it invalidates RN, invalidates the characteristic axioms of S4 and S5, and suggests a reconception of the nature of the contingency of actuality.

Bibliography

Adams, R., 'Theories of Actuality', *Noûs*, 8 (1974), 211–31.
____ 'Actualism and Thisness', *Synthese*, 49 (1981), 3–41.
Bennett, K., 'Two Axes of Actualism', *Philosophical Review*, 114 (2005), 297–326.
____ 'Proxy Actualism', *Philosophical Studies*, 129 (2006), 263–94.
Black, M., 'The Identity of Indiscernibles', *Mind*, 61 (1952), 153–64.

[83] Note that this restriction does not incorporate the restriction needed to accommodate the addition of $\mathcal{A}$, discussed above in n. 9.

Crossley, John, and Lloyd Humberstone, 'The Logic of "Actually" ', *Reports on Mathematical Logic*, 8 (1977), 11–29.

Deutsch, H., 'Contingency in Modal Logic', *Philosophical Studies*, 60 (1990), 89–102.

——— 'Logic for Contingent Beings', *Journal of Philosophical Research*, 19 (1994), 273–329.

Fine, K., 'Postscript: Prior on the Construction of Possible Worlds and Instants', in A. N. Prior and K. Fine, *Worlds, Times and Selves* (London: Duckworth, 1977), 116–68.

——— 'Modal Theory for Modal Logic, Part I—The *De Re/De Dicto* Distinction', *Journal of Philosophical Logic*, 7 (1978), 125–56.

——— 'Plantinga on the Reduction of Possibilist Discourse', in J. Tomberlin and P. van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel, 1985), 145–86.

Fitch, G. W., 'In Defense of Aristotelian Actualism', *Philosophical Perspectives*, 10 (1996), 53–71.

Grim, P., 'There is No Set of All Truths', *Analysis*, 46 (1986), 186–91.

Hazen, A., 'Actuality and Quantification', *Notre Dame Journal of Formal Logic*, 31 (1990), 498–508.

Hodes, Harold, 'Axioms for Actuality', *Notre Dame Journal of Formal Logic*, 31 (1984), 498–508.

Humberstone, L., 'Two-dimensional Adventures', *Philosophical Studies*, 118 (2004), 17–65.

Jager, T., 'An Actualist Semantics for Quantified Modal Logic', *Notre Dame Journal of Formal Logic*, 3 (1982), 335–49.

Jubien, M., 'Problems with Possible Worlds', in D. Austin (ed.), *Philosophical Analysis: A Defense by Example* (Dordrecht: Kluwer Academic Publishers, 1988), 299–322.

Kamp, Hans, 'Formal Properties of "Now" ', *Theoria*, 37 (1971), 227–73 [originally presented in 1967 at UCLA].

Kaplan, David, 'Demonstratives' (1977), in J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan* (Oxford: Oxford University Press, 1989), 481–564.

Kripke, S., 'Semantical Considerations on Modal Logic', *Acta Philosophica Fennica*, 16 (1963), 83–94.

——— '*Reference and Existence*', unpub. John Locke Lectures, 1973.

Lewis, D., 'General Semantics', *Synthese*, 22 (1970), 18–67.

——— 'Index, Context, and Content', in S. Kanger and S. Ohman (eds.), *Philosophy and Grammar* (Dordrecht: Reidel, 1980), 79–100.

——— *On the Plurality of Worlds* (Oxford: Basil Blackwell, 1986).

Linsky, B., and E. Zalta, 'In Defense of the Simplest Quantified Modal Logic', *Philosophical Perspectives*, 8 (1994), 431–58.

Marcus, R. B., 'Identity and Individuals in a Strict Functional Calculus of First Order', *Journal of Symbolic Logic*, 12 (1946), 3–23.

——— 'Modalities and Intensional Languages', *Synthese*, 13 (1961), 303–22.

——— 'Essentialism in Modal Logic', *Noûs*, 1 (1967), 90–6.

——— 'Dispensing with Possibilia', *Proceedings and Addresses of the American Philosophical Association*, 49 (1975), 39–51.

——— 'Possibilia and Possible Worlds', *Grazer Philosophische Studien*, 25–6 (1986), 107–33 [revised version printed in her *Modalities: Philosophical Essays* (New York: Oxford University Press), 190–213].

Meinong, A., 'On the Theory of Objects', in R. Chisholm (ed.), *Realism and the Background of Phenomenology* (Glencoe, Ill.: Free Press, 1960)1, 76–117 [German original first published 1904].

Menzel, C., 'Actualism, Ontological Commitment, and Possible Worlds Semantics', *Synthese*, 58 (1990), 355–89.

——— 'Singular Propositions and Modal Logic', *Philosophical Topics*, 21 (1993), 113–48.

——— 'Actualism', *The Stanford Encyclopedia of Philosophy* (Summer 2005 edn.), ed. E. Zalta <http://plato.stanford.edu/archives/sum2005/entries/actualism/>.

Nelson, M. (unpub. MS), 'Anti-Essentialism and *de re* Modality'.

——— and Zalta, E. (forthcoming), 'Bennett and "Proxy Actualism" ', *Philosophical Studies* (2007).

Parsons, T., *Nonexistent Objects* (New Haven, Conn.: Yale University Press, 1980).

Plantinga, A., *The Nature of Necessity* (Oxford: Oxford University Press, 1974).

——— 'Actualism and Possible Worlds', *Theoria*, 3 (1976), 139–60.

——— 'On Existentialism', *Philosophical Studies*, 44 (1983), 1–20.

——— (1985), 'Self-profile', in J. Tomberlin and P. van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel, 1985), 3–97.

Prior, A. N., *Past, Present, and Future* (Oxford: Clarendon Press, 1967).

——— *Papers on Time and Tense* (Oxford: Clarendon Press, 1968).

Quine, W. V., 'Notes on Existence and Necessity', *Journal of Philosophy*, 40 (1943), 113–17.

——— 'On the Problem of Interpreting Modal Logic', *Journal of Symbolic Logic*, 12 (1947), 43–8.

——— 'Reference and Modality', 2nd rev., in his *From a Logical Point of View* (New York: Harper and Row, 1980), 139–59 [original version published 1953].

Ray, G., 'An Ontology-free Modal Semantics', *Journal of Philosophical Logic*, 25 (1996), 333–61.

Salmon, N., 'Modal Paradox: Parts and Counterparts, Points and Counterpoints', *Midwest Studies in Philosophy*, 11 (1986), 75–120.

——— 'The Logic of What Might Have Been', *Philosophical Review*, 98 (1989), 3–34.

——— 'Nonexistence', *Noûs*, 32 (1998), 277–319.

Segerberg, Krister, 'Two-Dimensional Modal Logic', *Journal of Philosophical Logic*, 2 (1973), 77–96.

Stalnaker, R., *Inquiry* (Cambridge, Mass.: MIT Press, 1984).

Turner, J., 'Strong and Weak Possibility', *Philosophical Studies*, 125 (2005), 191–217.

van Inwagen, P., 'Creatures of Fiction', *American Philosophical Quarterly*, 14 (1977), 299–308.

Vlach, Frank, '"Now" and "Then": A Formal Study in the Logic of Tense Anaphora', 1973 Ph.D. diss., UCLA.

Walton, K., *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Cambridge, Mass.: Harvard University Press, 1990).

Williamson, T., 'Necessity and Existents', in A. O'Hear (ed.), *Logic, Thought, and Language* (Cambridge: Cambridge University Press, 2001), 233–51.

Zalta, Edward, 'Logical and Analytic Truths That are Not Necessary', *Journal of Philosophy*, 85 (1988), 57–74.

——— 'Natural Numbers and Natural Cardinals as Abstract Objects: A Partial Reconstruction of Frege's *Grundgesetze* in Object Theory', *Journal of Philosophical Logic*, 28 (1999), 619–60.

——— (unpub. MS), *Principia Metaphysica* <http://mally.stanford.edu/principia.pdf>.

4

Consciousness and Introspective Inaccuracy[1]

DERK PEREBOOM

A Kantian perspective on the nature of introspective awareness, I will contend, inspires a defense of a physicalist understanding of phenomenal states in response to the strongest arguments that have been raised against it. Immanuel Kant maintains that introspective representations—those of inner sense—are caused by the psychological states they represent and are wholly distinct from them, and they mediate the subject's awareness of those states, making it in a sense indirect. As a consequence, the subject may represent a psychological state as being a certain way, even though it is not really that way, or at least not as it is in itself.[2] I propose that the

[1] This paper was presented at Metaphysics, History, Ethics, a conference at Yale University in April 2005 in honor of Robert Adams, my dissertation advisor at UCLA. Thanks to Keith DeRose for his very fine comments at the session, and to the audience, especially Robert Adams, for high-quality questions and reflections. It was also presented in colloquia at the University of Auckland, the Australian National University, and the University of Alabama, and I'm grateful to audiences there for enlightening discussions. Thanks in addition to Kati Balog, Karen Bennett, Larry Jorgensen, Fiona Macpherson, Laurie Paul, Denis Robinson, David Kaplan, Adam Wager, Brian Weatherson, Sin yee Chan, and Hilary Kornblith for helpful comments and conversation. Special thanks are owed to Torin Alter, David Barnett, David Chalmers, David Christensen, Louis deRosset, Tyler Doggett, Mark Moyer, Nico Silins, and Daniel Stoljar for extensive and valuable commentary, discussion, and correspondence. Research on this article was facilitated by a generous Visiting Fellowship in the Centre for Consciousness of the Research School of Social Sciences at the Australian National University.

[2] Immanuel Kant, *Critique of Pure Reason*, trans. Paul Guyer and Allen Wood (Cambridge: Cambridge University Press, 1987), B152–4. Leibniz's views on perception provide another model for the idea that our introspective representations of our phenomenal states are inaccurate, as Robert Adams suggested to me. For instance, Leibniz claims: 'it does not cease to be true that at bottom confused thoughts are nothing other than a multitude of thoughts which are in themselves like the distinct, but which are so small that each separately does not excite our attention and cause itself to be distinguished. We can even say that there is at once a virtually infinite number of them contained in our sensations.' (G. W. Leibniz, *Die philosophischen Schriften*, ed. C. I. Gerhard, 7 vols. (Hildesheim: Olms, 1965), iv. 574–5; cf. Leibniz, *Discourse on Metaphysics*, 33, G iv. 458–9). Contained in our sensations is a virtually infinite number of thoughts, so 'small' that they are not consciously distinguished. G. H. R. Parkinson

possibility of this sort of inaccuracy yields a significant challenge to Frank Jackson's knowledge argument,[3] and that it provides the physicalist with a response to those, like Joseph Levine and Robert Adams, who suggest that there is an explanatory gap between the physical and the phenomenal that we do not know how to close.[4]

1. The Qualitative Accuracy Intuition

In Jackson's story, Mary has lived her entire life in a room that displays only various shades of black, white, and gray.[5] She acquires information about the world outside, and also about the physical nature of the human being, by means of a black and white television monitor. By watching television programs Mary eventually comes to have knowledge of all of the physical information there is about the nature of the human being. (This complete physical knowledge might be conceived as complete microphysical knowledge, or as complete knowledge of any entity that is uncontroversially physical, or else as exhaustive factual knowledge of every entity that is wholly physically constituted—each of these proposals might have its advantages and disadvantages.) But even if she knows all of this, Jackson contends, there is much she will not know about human experience. She will not know, for example, what it is like visually to experience a ripe red tomato, and, in particular, she lacks knowledge of

points out that in the late 1670s and beyond Leibniz held that it is impossible for us to reach genuinely primitive concepts in our analysis of sensations (Parkinson, *Leibniz, Logical Papers* (Oxford: Clarendon Press, 1966), pp. xxvii–xxviii, 51–2; for a discussion of this point, see my 'Kant's Amphiboly', *Archiv für Geschichte der Philosophie*, 73 (1991), 50–70). But, as Nico Silins remarked, then our sensations would be a certain way even though they are not introspectively represented by us as being that way, while the stronger distinctively Kantian claim I am singling out is that we introspectively represent sensations (for example) to be a certain way, even though they are not that way, at least as they are in themselves. In Silins's helpful terminology, Leibniz's claim is that introspective representation is merely silent about certain features of sensations, while Kant's idea is that it is in one respect mistaken about them.

[3] I think that David Chalmers's zombie argument is also vulnerable to this strategy, although I defer development of this claim to another occasion (but see n. 31).

[4] Joseph Levine, 'Materialism and Qualia: The Explanatory Gap', *Pacific Philosophical Quarterly*, 64 (1983), 354–61, and id., *Purple Haze: The Puzzle of Consciousness* (Oxford: Oxford University Press, 2001); Robert Adams, 'Flavors, Colors and God', in Adams, *The Virtue of Faith* (Oxford: Oxford University Press, 1987), 243–62.

[5] Frank Jackson, in 'Epiphenomenal Qualia', *Philosophical Quarterly*, 32 (1980), 127–36, and in 'What Mary Didn't Know', *Journal of Philosophy*, 83 (1986), 291–5; cf. Thomas Nagel, 'What is it Like to Be a Bat?' *Philosophical Review*, 83 (1974), 435–50.

what it is like to see red. When she leaves the room and sees a red tomato, she comes to know for the first time—she *learns*—what it is like to see red. She gains knowledge, for the first time, of a particular *phenomenal property*, or of a mental state that has this property—a *phenomenal state*.[6] Thus there are facts about phenomenal states that are not physical facts, and thus phenomenal states are not completely physical. The core intuition underlying the knowledge argument is that if someone who possesses complete physical knowledge does not *thereby* know some fact about a phenomenal state, then that fact cannot be physical, and, moreover, the phenomenal state cannot be completely physical.[7]

Now consider the 'old fact/new guise' response to the knowledge argument—which I do not endorse, but the reply I develop can be understood as a successor to it.[8] According to this kind of response, Mary, when she is still in the room, does indeed know every fact about phenomenal states by virtue of her exhaustive physical knowledge, while what she is missing are only ways of introspectively representing those states, or, as I will put it, *introspective modes of presentation* of those states.[9]

[6] David Chalmers characterizes phenomenal properties as those that 'type mental states by what it is like to have them', 'The Content and Epistemology of Phenomenal Belief', in Q. Smith and A. Jokic (eds.), *Consciousness: New Philosophical Perspectives* (Oxford, 2003). The 'what it is like to have them' locution should perhaps be taken as a means of signaling to an audience what to look for as instances of phenomenal properties, which can then serve as paradigms, and not so much as a thorough descriptive characterization of this type of property.

[7] One reply to the argument is that the reason that pre-emergence Mary lacks knowledge of phenomenal states is just that she is missing phenomenal concepts. In response, Daniel Stoljar strengthens the argument by specifying that pre-emergence Mary possesses all the phenomenal concepts, while she nevertheless lacks knowledge of how correct applications of phenomenal concepts are correlated with physical states; 'Physicalism and Phenomenal Concepts', *Mind and Language* 20 (2005); Stoljar tells a plausible story as to how Mary might come to fit this description (after acquiring the phenomenal concepts Mary suffers selective amnesia); David Chalmers makes a similar point in 'The Two-Dimensional Argument Against Materialism', in *The Character of Consciousness* (unpublished MS). The resulting argument has a somewhat different focus. What I say in section 7 in reply to Adams is also a response to this argument.

[8] Proponents of the 'old fact/new guise' response include Terence Horgan, 'Jackson on Physical Information and Qualia', *Philosophical Quarterly*, 32 (1984), 147–52; Paul M. Churchland, 'Reduction, Qualia and the Direct Introspection of Brain States', *Journal of Philosophy*, 82 (1985), 8–28; Robert Van Gulick, 'Physicalism and the Subjectivity of the Mental', *Philosophical Topics*, 13 (1985), 51–70; Michael Tye, 'The Subjective Qualities of Experience', *Mind*, 95 (1986), 1–17; Brian Loar, 'Phenomenal States', *Philosophical Perspectives*, 4; James Tomberlin (ed.), *Action Theory and Philosophy of Mind* (Atascadero, Calif.: Ridgeview, 1990), 81–108; William G. Lycan, 'What is the "Subjectivity" of the Mental?', *Philosophical Perspectives*, 4 (1990), 109–30.

[9] I use the Fregean term 'mode of presentation' as a convenient nominalization, without intending the full Fregean theory. The claims made in this paper can generally be made in more neutral terms, or in terms of other theories of cognition and language.

When she leaves the room and sees the red tomato, she comes to represent a phenomenal state, about which she already knew everything, by an introspective mode of presentation, with which she had never represented that phenomenal state while she was in the room. In this way, the appearance of Mary's coming to know a new fact can be explained without granting that she acquires new knowledge.

This sort of reply might be illustrated by various analogies. According to William Lycan, the difference between the introspective and the physical representations is akin to the difference between my use of 'I' and your use of 'you' to represent me in the representation of some fact about me.[10] For example, consider:

(1) 'I weigh 195 pounds' (asserted by me)
(2) 'You weigh 195 pounds' (asserted by you).

You cannot represent the fact that I weigh 195 pounds by 'I weigh 195 pounds', whereas I can represent this fact by means of that sentence. But suppose that you have knowledge of this fact, and represent it by 'You weigh 195 pounds'. Then there is no fact of which I have knowledge but you don't; the only fact to be known here is that DP weighs 195 pounds, and we both know it.

Although some find analogies of this sort sufficient to dislodge the knowledge argument, its proponents remain unconvinced.[11] To advance the debate, we need to explore why the argument has this residual force. Are there features of Mary's epistemic situation disanalogous with Lycan's example that might explain this force? Phenomenal states have characteristic phenomenal properties, and it is intuitive, for some, at least, that:

(i) Both the physical and introspective modes of presentation represent these phenomenal properties as having a qualitative nature, and the specific qualitative nature that the introspective mode of presentation represents a phenomenal property as having is not included in the

[10] Lycan, 'What is the "Subjectivity" of the Mental?'

[11] A sophisticated reply along these lines is provided by John Perry, *Knowledge, Possibility, and Consciousness* (Cambridge, Mass.: MIT Press, 2001); see also John Hawthorne, 'Advice for Physicalists', *Philosophical Studies*, 109 (2002), 53–74; for someone who is not convinced by these sorts of accounts, see Chalmers, 'Imagination, Indexicality and Intensions', *Philosophy and Phenomenological Research*, 68 (2004), 182–90.

qualitative nature the physical mode of presentation represents it as having.[12]

It is also intuitive—again, for some—that:

(ii) The introspective mode of presentation *accurately represents* the qualitative nature of the phenomenal property. That is, the introspective mode of presentation represents the phenomenal property as having a specific qualitative nature, and the attribution of this nature to the phenomenal property is correct.

There is no uncontroversial way to characterize the qualitative nature that introspective modes of presentation represent phenomenal properties as having. One option, inspired by John Locke, is to characterize this nature by way of resemblance to modes of presentation. Thus, in our example, we might say that Mary's introspective representation of her phenomenal-red sensation presents that sensation in a what-it-is-like-to-sense-red way, and it is intuitive that a qualitative nature that resembles this what-it-is-like mode of presentation is correctly attributed to the phenomenal property.[13] Or, in deference to concerns about the cogency of such resemblance characterizations, one might say simply that the qualitative nature of the phenomenal property is as it is introspectively represented.[14]

[12] Joseph Levine, in *Purple Haze*, accounts for the existence of the explanatory gap partly by the fact that 'modes of presentation whereby we come into cognitive content with qualia are substantive and determinate' (p. 8) and that 'there is real content to our idea of a quale' (p. 84). In what I am saying here I aim to explicate these kinds of intuitions; cf. Alex Byrne, 'Review of Purple Haze', *Philosophical Review*, 111 (2002), 594–7.

[13] Accepting a resemblance claim of this sort does not amount to endorsing a discredited *resemblance theory of representation*, as is sometimes suggested. For in accepting that the phenomenal property accurately attributed to the state resembles the introspective mode of presentation, one is not also accepting a resemblance account of how it is that the mode of presentation represents the phenomenal property. By analogy, one does not have to accept a resemblance account of photographic representation in order to accept the claim that photographs can resemble what they represent.

[14] Note that the accuracy claim (ii) is weaker than what David Lewis calls *revelation*, which is: 'when I have an experience with quale Q, the knowledge I thereby gain reveals the essence of Q: a property of Q such that, necessarily, Q has it and nothing else does'; ('Should a Materialist Believe in Qualia', *Australian Journal of Philosophy*, 73 (1995), 140–4, repr. in Lewis's *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press, 1999), 325–31 at 328). The accuracy claim is not that the essence of the quale is revealed in an (introspective) experience of the quale, but rather that the quale really has the qualitative nature this experience represents it as having. This is consistent with this experience not representing the complete essence of the quale.

Given these claims about what is intuitive, an advocate of the knowledge argument can account for its residual force in the following way. When Mary leaves the room and sees the tomato, she comes to believe that

(T) Phenomenal redness has qualitative nature Q.

Qualitative nature Q is accurately represented introspectively by way of the *what-it-is-like-to-sense-red* introspective mode of presentation. But on the physicalist hypothesis, every truth about the qualitative nature that an introspective mode of presentation accurately represents a phenomenal property as having would need to be derivable from a proposition detailing features that physical modes of presentation represent the world as having. However, (T) is not derivable from such a proposition. So not every truth about the qualitative nature that the introspective mode of presentation accurately represents the phenomenal property as having is so derivable. So the physicalist hypothesis is false.[15]

One might challenge this version of the knowledge argument at various points. In particular, one might take issue with one or both of the claims about what is intuitive just listed. The one I will dispute is (ii), the claim about the accuracy of introspective representation. I will leave (i) as common ground, and continue the discussion with the supposition that (i) is in fact true. On (ii), in my view it is an epistemic possibility of a certain sort that introspective modes of presentation represent phenomenal properties as having certain specific qualitative natures, while they do not in fact have them, and that introspective representation is in this sense inaccurate. Of the many notions of epistemic possibility, the sense I here have in mind is: possible given what we human beings now rationally believe. (The relevant 'we' in this case are perhaps those who have thought carefully about these philosophical issues.) For this sense of epistemic possibility, I will use the term 'open possibility'.

[15] On an alternative version of the 'old fact/new guise' response, phenomenal modes of presentation should not be taken to represent phenomenal properties as having a qualitative nature at all. Rather, they are just devices for securing reference to phenomenal properties, analogous to demonstratives. There would then be no good reason to think that the physical and phenomenal modes of presentation of phenomenal properties are not co-referential. To my mind, this sort of response to the knowledge argument is weakened in its effectiveness by the plausibility of the claim that phenomenal modes of presentation represent phenomenal properties as having a qualitative nature.

2. Is Qualitative Inaccuracy a Serious Open Possibility?

My contention, then, is that, given the supposition of (i), it is an open possibility that introspective representation is inaccurate in that it represents phenomenal properties as having qualitative natures they do not in fact have. For instance, it is an open possibility that upon seeing the red tomato Mary introspectively represents the qualitative nature of phenomenal redness in the *what-it-is-like-to-sense-red* way, and her representing it in this way attributes to it a qualitative nature that it actually lacks.

The notion that there might be such a discrepancy between the real nature of phenomenal properties and the qualitative natures we introspectively represent them as having is consistent with certain claims about the correctness of introspective representation. For example, even if introspective representation inaccurately represents phenomenal properties in the sense just outlined, still it may be that a belief *that I am in* a phenomenal state characterized by a certain phenomenal property, a belief that is formed on the basis of an introspective representation (perhaps a belief that does not feature a linguistic term for the phenomenal state), is generated by a mechanism that is very reliable. So in general there might be no discrepancy between which phenomenal states I introspectively represent myself as being in and those I am actually in—introspective representation might sort phenomenal states and properties quite accurately—while at the same time phenomenal properties lack the qualitative natures we introspectively represent them as having.

In this view, a type of representation might successfully secure a referent by, for example, having instances that are caused by this referent, and yet misrepresent this referent by representing it as having a property that it really lacks. Locke's conception of sensory secondary quality representation provides an analogy. He thinks that these representations do indeed secure their referents causally, while they nevertheless misrepresent external objects in a certain respect:

> Ideas of primary qualities are resemblances; of secondary, not. From which I think it easy to draw the observation that the ideas of primary qualities are resemblances of them and their patterns do really exist in the bodies themselves, but the ideas produced in us by these secondary qualities have no resemblance

of them at all. There is nothing like our ideas existing in the bodies themselves.[16]

On a plausible interpretation of this view, our ordinary tactile representations of temperature represent the ambient air, or the icicles above the door, or the coffee one is drinking, as having certain features, while those features are incorrectly attributed to those things. On a warm day, we have a particular sort of tactile temperature representation of the ambient air, which represents the air as having a certain feature—put in Lockean terms, as having a quality that resembles the sensory temperature idea. However, if Locke is right, that 'primitive' quality is incorrectly attributed to the air. William Alston might well be endorsing a view of this type when he says: 'when I look at a shirt and take it to be red, when I feel a fabric and recognize it as very smooth, when I hear a bell ringing and recognize it as giving out a typical bell-like sound, I attribute to the perceived objects qualities that they do not, in strictness, bear.'[17]

Locke's contentions about sensory representations of secondary qualities are controversial. Some would dispute the claim that there is any sense in which our ordinary visual color representations generally misrepresent, for the reason that what a type of representation represents is determined solely by the typical cause of its instances. Claims of this last sort have often been disputed by way of devices such as inverted spectrum thought experiments, in which what is represented in the external world is held fixed, while the phenomenal content of the representation varies. Familiarly, there is widespread disagreement about the force of the attendant argument. Nevertheless, I will make use of the secondary quality analogy, assuming Locke's position that, for example, our ordinary visual color sensations represent physical objects as having qualitative features that are incorrectly attributed to them. A more localized example is that, as Descartes pointed out, from a certain distance we visually represent square towers as round,

[16] John Locke, *An Essay Concerning Human Understanding* (Oxford: Oxford University Press, 1975), II. viii. For a sympathetic exposition of Locke's position on this issue, see Michael Jacovides, 'Locke's Resemblance Theses', *Philosophical Review*, 108 (1999), 461–96.

[17] William Alston, 'Mystical and Perceptual Awareness of God', in *The Blackwell Guide to Philosophy of Religion*, ed. William E. Mann (Oxford: Blackwell Publishers, 2004), 211. Alston continues: 'No doubt, I could, in principle, restrict myself to beliefs that do not suffer falsity in this respect. I could, instead of taking the shirt to be red, take it to have primary qualities of such a sort that when it is seen under these conditions by a human being with normal vision, it will appear to have the color I call red. But that requires considerable reflection of the sort we do not typically engage in when perceiving things.'

while the property of roundness is incorrectly attributed to the tower.[18] Another is that we visually represent the lengths of the Müller–Lyer pair of lines as different, while they are in fact the same. It is the open possibility of an analogous disparity between how phenomenal properties are represented introspectively and their real nature that would generate the physicalist response.

In the case of our visual color representations, their specific causal nature is plausibly what allows for a disparity of this kind. By the standard causal theory of such representations, they are typically caused by external objects (perhaps by their property-instances) and are distinct from them, and they mediate the subject's representation of the object, rendering it indirect. The mediation of these representations can result in a disparity between the features objects appear to have and those they really have, and so representation may in certain respects be inaccurate. For instance, they might represent the color of an object as having a specific qualitative nature, while the attribution of this nature to the property is incorrect. It is an open possibility that our introspective representation of phenomenal properties is similarly causal, whereupon a guarantee of the accuracy of how introspection represents these properties would be precluded.

Non-causal theories of introspective representation are also contenders. One might, with Franz Brentano, endorse a *self-presentation* view, contending that a token sensation of green is on the one hand a sensation of green, but at the same time that very sensation is also an experience of itself.[19] Alternatively expressed, the idea is that besides representing to the subject the property of being green, this sensation also simply presents itself to her without the mediation of a (further) representation of it. So in one kind of case—when a sensation is an experience of itself—representation of something occurs without causal mediation. Representation is instead reflexive and non-causal.[20] Perhaps a self-presentation view meshes with

[18] René Descartes, *Meditations on First Philosophy*, in *The Philosophical Writings of Descartes*, ii, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press, 1984); Meditation 6.

[19] Franz Brentano, *Psychology from an Empirical Standpoint*, trans. A. C. Rancurello, D. B. Terrell, and L. L. McAlister (London: Routledge & Kegan Paul, 1973), 153–4. Uriah Kriegel develops this position in forthcoming work.

[20] Christopher Hill and Brian McLaughlin explain this position as follows: 'Sensory states are self-presenting states: we experience them, but we do not have sensory experiences of them. We experience them by *being in* them. Sensory concepts are recognitional concepts: deploying such concepts, we can introspectively recognize when we are in sensory states simply by focusing our attention directly on

certain of our ordinary intuitions about our consciousness of sensation. To my mind it is also an open possibility.

Or with David Chalmers one might advocate a *constitution* view for (pure) phenomenal concepts. He says: 'one might say very loosely that the referent of the concept is somehow present inside the concept's sense, in a way much stronger than in the usual cases of "direct reference" . . . in the phenomenal case, the epistemic content itself seems to be constituted by the referent.'[21] Here again a phenomenal property would be represented without causal mediation. Perhaps this position is not at odds with the self-presentation view, but one might envision it being developed so that it is clearly different.

It may seem strongly intuitive that phenomenal properties are introspectively represented in an intimate way that guarantees that their qualitative nature is represented accurately. The color of a physical object might not be accurately represented by way of our ordinary sensory representations, but how could the qualitative nature of pleasure, or the qualitative nature of one's visual sensation of red, not be as it is introspectively represented? But although this sort of discrepancy might be at odds with strong intuitions, still its being an open possibility is forced on us by the prospect that introspection might be causal on analogy with visual color representation.

Moreover, the way we naturally come to think that there are mediating representations in the case of external sensation is that here we fairly easily and frequently become aware of a discrepancy between the real nature of what is represented and the way it is represented as being. The car appears to have a different color under the sodium vapor lights than it does in natural light, but it is clear that nothing about the car itself has changed, and so we come to believe that there is a discrepancy between the car's real color and the way it is visually represented under the unusual lighting conditions. But for introspection of phenomenal properties, awareness of such discrepancies would not readily arise, supposing they existed. Perhaps

them. Matters are of course quite different in the case of perceptual and theoretical concepts. An agent's access to the phenomena that he or she perceives is always indirect: it always occurs via an experience of the perceived phenomena that is not identical with the perceived phenomena, but rather caused by it.' 'There are Fewer Things in Reality than are Dreamt of in Chalmers's Philosophy', *Philosophy and Phenomenological Research*, 59 (1999), 448.

[21] Chalmers, 'The Content and Epistemology of Phenomenal Belief', 13–14.

they *sometimes* arise: Christopher Hill cites the following case, presented by Rogers Albritton in seminar:

> The case involves a college student who is being initiated into a fraternity. He is shown a razor, and is then blindfolded and told that the razor will be drawn across his throat. When he feels a sensation he cries out: he believes for a split second that he is in pain. However, after contemplating the sensation for a moment, he comes to feel that it is actually an experience of some other kind. It is, he decides, a sensation of cold. And this belief is confirmed when, a bit later, the blindfold is removed and he is shown that his throat is in contact with an icicle rather than a razor.[22]

There are a number of ways to analyze this example, but one possibility is that in his introspective awareness the fraternity pledge misrepresents the qualitative features of the sensation of cold he actually has as qualitative features of pain. But this would be a controversial analysis.[23] Here is a possible example of a pain sensation that is introspectively misrepresented as a sensation of cold. My daughter recently required a Novocain shot at the dentist. Rather than simply showing her the needle in advance, and then giving her the injection, the dentist hid the needle from her, and told her that he would be dropping bits of cold water into her mouth. She didn't flinch. When I asked her afterward whether the experience was unpleasant, she said that she didn't like the drops of water much, but they didn't hurt. Here we might want to say that the dentist's suggestion, together with his hiding the needle, kept her from introspectively representing the qualitative features of the pain state she was actually in as qualitative features of pain, while instead she misrepresented those features as qualitative features of a sensation of cold. This, again, is a controversial analysis. But my point here is that if such examples of misrepresentation occur at all, it is only infrequently. They are too unusual to give rise to a vivid sense that when

[22] Hill, *Sensations: A Defense of Type Materialism* (Cambridge: Cambridge University Press, 1991), 128–9.

[23] Hill takes this example to provide evidence that we make *errors of judgment* in our introspection-based beliefs about sensation, where errors of judgment 'are usually due either to some form of inattention or to the influence of expectation upon judgment'. He differentiates between errors of judgment and *errors of ignorance*, which occur 'when beliefs are based on appearances that fail to do justice to the entities to which the beliefs refer'. Hill claims 'we are perforce innocent of committing errors of ignorance in forming beliefs about our own sensations' (*Sensations*, 127–8). The open possibility I am envisioning would have us making errors of ignorance in our introspection-based beliefs about phenomenal properties, since such beliefs would be based on appearances that fail to do justice to the real qualitative nature of those properties.

we introspectively represent phenomenal properties, we might represent them as having natures that they actually lack.

In addition, in the external case, we have readily available ways of checking the object represented that are independent of the representation under scrutiny, while in the introspective case such a capacity is at best very limited. One might have a closer look at Descartes's tower in order to check whether one's visual representation of its shape as round was accurate, or measure the Müller–Lyer lines to determine whether one's visual representation of them as having different lengths was correct. But analogous ways of checking introspected phenomenal properties are not similarly available, if we have them at all.

These observations help explain our resistance to the idea that introspection of phenomenal properties features mediating representations. Given that awareness of a discrepancy between the real nature of the phenomenal property and the qualitative nature we introspectively represent it as having would seldom, if at all, arise, and given the scarcity of means of checking the accuracy of such representations, there would be little if any noticeable difference between an introspective experience in which we represented phenomenal properties causally and such discrepancies did exist as a result, and one in which phenomenal properties (or the sensations or states of which they are properties) were self-presenting, or one for which the constitution view held, and these discrepancies did not exist. Hence, what we do and do not notice about introspective experience, all by itself, does not adjudicate the controversy about how phenomenal properties are represented.[24]

While the self-presentation and constitution views are candidates for the correct account of introspective phenomenal representation, a causal theory is also a very serious contender. It is then also a serious open possibility that we introspectively represent phenomenal properties as having qualitative natures that they lack.

[24] Stephen Wykstra proposes the following plausible *condition of reasonable epistemic access*: 'On the basis of cognized situation s, human H is entitled to claim "It appears that p" only if it is reasonable to believe that, given her cognitive faculties and the use she has made of them, if p were not the case, s would likely be different than it is in some way discernible by her;' ('The Humean Obstacle to Evidential Arguments from Suffering; On Avoiding the Evils of "Appearance"', *International Journal for Philosophy of Religion*, 16 (1984), 73–93, repr. in M. M. Adams and R. M. Adams (eds.), *The Problem of Evil*, 138–60 at 152). By Wykstra's criterion, we would not be justified in claiming that it appears that introspective representation of phenomenal properties is non-causal.

Moreover, the self-presentation view of phenomenal representations does not obviously preclude such qualitative inaccuracy. Self-presenting sentences can misrepresent in some respect (while being accurate in another). 'This German sentence has six words,' for instance, represents itself as having a feature it lacks—as being a German sentence (while it accurately represents itself as having six words).[25] Perhaps, then, nothing we know rules out the possibility that self-representing phenomenal states are qualitatively inaccurate in the sense we have been discussing. It may also be that the constitution view does not preclude such qualitative inaccuracy—we would need to be told more about how it works. Still, I think that the stronger case for my position can be made by analogy with secondary-quality representation, and here it is reasonable to believe that qualitative inaccuracy is due to causal mediation, and not to the sort of problem that arises in the case of misrepresenting self-presenting sentences. Nonetheless, if the requisite kind of qualitative inaccuracy for introspective representations of phenomenal properties is an open possibility even if phenomenal states are self-presenting, then my case will be stronger.

It should be noted that in the proposed open possibility the specified kind of inaccuracy is universal—it is a feature of all human introspective representation of the qualitative nature of phenomenal properties; and it is in a significant respect extensive—the qualitative natures that phenomenal properties are represented as having they altogether lack. In these respects, the inaccuracy at issue differs from that of the sort we sometimes make when we visually represent the lengths of pairs of lines, or when we visually represent the shapes of objects from a significant distance.[26] One might contend that the universality and extensiveness of the proposed inaccuracy provides reason to believe that the open possibility under consideration is very unlikely to be actual. However, on the Lockean theory of our sensory representations of secondary qualities, which is not an implausible theory, the inaccuracy of these representations is similarly universal and extensive. To my mind, this analogy yields significant reason to believe that the proposed open possibility is not unlikely to be actual. Moreover,

[25] Mark Moyer and Brian Weatherson each made this point about self-representing sentences and suggested that the possibility of misrepresenting self-representation would strengthen the argument. Louis deRosset provided the example of a self-presenting sentence that is accurate in one respect and inaccurate in another.

[26] Thanks to Louis deRosset for this point.

as I remarked earlier, this open possibility is compatible with introspective representation reliably generating true beliefs about which phenomenal state the subject is in. So this possibility can preserve the accuracy of many introspectively-based beliefs about phenomenal states, and the extensiveness of introspective misrepresentation of phenomenal states that it involves is actually quite limited.

Some might still have the sense that the qualitative inaccuracy for our introspective representations of phenomenal properties is implausibly as universal and extensive as required to yield a promising response to the knowledge argument. But it may be that all of the developed positions on the metaphysics of consciousness have implausible features that should make for at least some resistance to belief. As a case in point, Karen Bennett argues that the traditional dualist position has such an implausibility: it accepts that there exist a fairly large number of psychophysical laws that are brute in the sense that there is no explanation as to why they hold, or for which the explanation we can envision is arbitrary divine preference.[27] (Locke suggests the divine preference explanation, and Adams endorses it.[28]) With this in mind, should one be less resistant to the traditional dualist view than to our qualitative inaccuracy claim? Alternative physicalist hypotheses also have elements that are to at least some degree implausible, as their opponents, such as Chalmers, have contended.[29] Should one be less resistant to these physicalist views than to the qualitative inaccuracy claim? There are a number of central philosophical issues for which all defended positions are in some key respect implausible—free will and moral responsibility is a case in point. For such issues, it is not sufficient to dislodge a position for it to be in some way implausible, and the metaphysics of consciousness might well be such an issue.

3. A Response to the Knowledge Argument

While Mary is in the room, she does not represent phenomenal states in the characteristic introspective way, and she does not seem to have

[27] Karen Bennett, 'Why I Am Not a Dualist' (MS).

[28] Locke, *Essay Concerning Human Understanding*, IV. iii, 6, 28–9; Adams, 'Flavors, Colors and God', 241–51.

[29] See, for example, Chalmers, 'Consciousness and its Place in Nature'.

the information required to represent the complete real natures of these phenomenal states by deriving it from what she knows. But it is a serious open possibility that by virtue of her physical knowledge she can nevertheless accurately represent the complete real natures of these states. For phenomenal properties of these states, in particular, might not have the qualitative natures they are introspectively represented as having. Instead, the natures of these properties might accurately be represented by way of Mary's physical knowledge. For from her physical knowledge she might then be able to derive every truth about the natures of phenomenal states.[30]

How exactly does this story yield a response to the knowledge argument? First of all, let's focus on which true beliefs, and not which knowledge, Mary has before and after she leaves the room. For what is germane to the knowledge argument when it comes to Mary's states while she is in the room is just that they are true beliefs, and thus we can set aside the complex concerns that are specific to knowledge. The key issue is whether upon leaving the room and seeing the red tomato Mary acquires a true belief that is new in the sense that she neither had it while she was in the room, nor was it derivable from the true beliefs she had then. We'll initially suppose that our open possibility is actually realized, but subsequently we'll discharge this supposition, and instead think of the open possibility as a hypothesis about how things might turn out to be. We'll then ask the crucial question: do we theorists now have good reason to believe that Mary has not acquired a new true belief?

[30] Chalmers (in 'Phenomenal Concepts and the Explanatory Gap', Torin Alter and Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (Oxford University Press, 2006)) points out that on the sort of view advocated by Loar, Levine, and others—the 'phenomenal concept strategy'—it is maintained that zombies are ideally, positively, primarily conceivable, while our having phenomenal concepts has a physical explanation. These are two key features of what Chalmers calls Type-B materialism, a widely held view. He argues that this position is unstable. I suspect that he is right about this, and that in the last analysis, physicalism requires denying the ideal, positive, primary conceivability of zombies, for then it would avoid the tension between affirming the conceivability of zombies, which has the consequence that phenomenal properties are in the crucial sense not physically explainable, and claiming that our having phenomenal concepts is physically explainable. Another key feature of the phenomenal concept strategy is its claim that '$P \rightarrow Q$' is a posteriori, and Daniel Stoljar ('Physicalism and Phenomenal Concepts') argues that it cannot adequately explain how this can be so. The response I've developed here does not require that this conditional is a posteriori, and thus it avoids the issue Stoljar highlights.

So, under the supposition that the open possibility is in fact realized, how should we describe what happens when Mary leaves the room and sees the red tomato? We imagine her coming to believe:

(T) Phenomenal redness has qualitative nature Q.

Consider first the initially plausible proposal (i) that in (T), the term 'Q' refers to a property accurately represented by the what-it-is-like-to-sense-red mode of presentation. On our open possibility, phenomenal redness has no such property, so what she comes to believe will be false. Thus she does not acquire a new true belief. Next, consider the perhaps initially less plausible proposal (ii) that the term 'Q' in (T) refers to a physical property that appears to Mary in the what-it-is-like-to-sense-red way, but whose qualitative nature is misrepresented by this mode of presentation. Under this interpretation, we can suppose that (T) is true, but since while in the room she already believed the truth expressed by (T), or was able to derive it from the true beliefs she already had, she also does not acquire a new true belief.

Now consider the open possibility just as a hypothesis about how things might turn out to be. Does this give us, as theorists, a good reason to believe that Mary hasn't acquired a new true belief? In my estimation, the open possibility is serious enough to provide us with such a reason. In fact, my sense is that this possibility is sufficiently serious to preclude rational belief that Mary does acquire a new true belief, and herein lies the challenge to the knowledge argument. How high would our rational credence be that she acquires a new true belief? I'll leave this to the reader to decide, but I have more to offer by way of an argument that it is not especially high.[31] But note that this is consistent with this rational credence nevertheless being quite substantial. As I conceive it, the seriousness of the open possibility does not provide us with good reason to believe that the anti-physicalist consideration raised by the knowledge argument has no force, but it does yield a significant reason to believe that this consideration fails to establish that physicalism is false.[32]

[31] Thanks to Nico Silins for discussion about this section.

[32] David Chalmers's zombie argument can be challenged in a similar way (Chalmers, *The Conscious Mind* (Oxford: Oxford University Press, 1996); id., 'Consciousness and its Place in Nature', *Blackwell Guide to the Philosophy of Mind* (Oxford: Blackwell, 2002), repr. in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings* (Oxford: Oxford University Press, 2002), 247–72; id., 'Does Conceivability Entail Possibility', in Tamar Gendler and John Hawthorne (eds.), *Conceivability and Possibility* (Oxford: Oxford University Press, 2002), 145–200). Let 'P' be a statement that details the

4. Hasn't the Problem for Physicalism Been Shifted to Modes of Presentation?

It may now seem that the problem for a physicalist explanation of consciousness has shifted from accounting for phenomenal states and their properties to accounting for their introspective modes of presentation. Supposing that the way phenomenal states are represented introspectively might be inaccurate in the way specified, and that Mary can derive every truth about the real nature of phenomenal states from her complete physical knowledge, the pressing issue is now to assess whether these introspective modes of presentation, or, perhaps more precisely, states featuring these modes of presentation, could have a physical account.[33] In fact, Chalmers develops this point as an objection to the old fact/new guise strategy. He contends that even if what Mary gains when she leaves the room

is only knowledge of an old fact under a different mode of presentation—then there must be some truly novel fact that she gains knowledge of. In particular, she must come to know a new fact involving that mode of presentation. Given that she already knew all the physical facts, it follows that materialism is false. The physical facts are in no sense exhaustive.[34]

physical truth about the actual world, and 'Q' an arbitrarily selected actual phenomenal truth. The first premise is: 'P and ~Q' is ideally, positively, primarily conceivable, from which we are to derive that 'P and ~Q' is metaphysically possible. This derivation requires the crucial assumption that we have representations that accurately represent the qualitative nature of phenomenal properties and the states that have them. More specifically, it requires that if it is not the case that a state has a property that is accurately represented by an introspective mode of presentation of the what-it-is-like-to-sense-x variety, it is not the case that the phenomenal property represented by this mode of presentation is instantiated by the state. For example, consider the phenomenal concept 'R', a concept of phenomenal property R, which represents R by way of the introspective mode of presentation *what-it-is-like-to-sense-red*. The argument demands that if it is not the case that a state has a property whose qualitative nature is accurately represented by the what-it-is like-to-sense-red introspective mode of presentation, then that state does not instantiate phenomenal property R. But, for this last conclusion to be established, it would have to be shown that it is not an open possibility for the phenomenal concept 'R' to be qualitatively inaccurate. For if this representation of R is indeed qualitatively inaccurate in this respect, then the fact that R *as represented in the what-it-is-like-to-sense-red way* is not derivable from 'P' fails to show that a description of the *real nature* of R is not derivable from 'P.' Thus it also does not show that 'Q', our selected truth about R, is not derivable from 'P'. Then, what we were thinking of as the zombie-world might not be one in which 'Q' is false after all. So, it seems that until the open possibility of qualitative inaccuracy has been closed off, the soundness of the zombie argument is in doubt.

[33] I consider this objection in 'Bats, Brain Scientists, and the Limitations of Introspection', *Philosophy and Phenomenological Research*, 54 (1994), 323–6.

[34] Chalmers, *The Conscious Mind*, 142.

Torin Alter raises a similar objection to my earlier account:

> How color sensations appear from the first-person perspective is itself a fact about them. Therefore, if when Mary is released she learns how they appear from the first-person perspective, then she learns a new fact about them. This is true regardless of whether this appearance accurately reflects the way they really are.[35]

So would Mary, by virtue of her physical knowledge, be able to derive every truth about the introspective mode of presentation of her phenomenal-red sensation—call it MP_R, or every truth about representational states that have MP_R as a component? In response, there is no less reason to think that a causal theory is true for introspective representations of introspective modes of presentation, or for introspective representations of states that have introspective modes of presentation as a component, than it is for phenomenal states themselves. Consequently, it is also a serious open possibility that our introspective representation of a state that has MP_R represents that state as having a qualitative feature it really lacks. Then, despite how it is introspectively represented, it might be that Mary can derive every truth about the real nature of a state that has MP_R as a component from her physical knowledge. So even though pre-emergence Mary will not have an introspective representation of this state, it is an open possibility that while she is in the room she can come to know every truth about it. Furthermore, a reply of this kind can be made to count against any iteration of this sort of objection.[36]

The success of the knowledge argument depends on there being phenomenal states or aspects of phenomenal states whose qualitative natures are as they are introspectively represented. Alter and Chalmers are right to argue that the 'old fact/new guise' response to the knowledge argument typically transfers the physicalism-challenging feature from the phenomenal state to the introspective mode of presentation. However, the qualitative inaccuracy move can be reiterated for introspective modes of presentation. Notice that the view that results from this is no longer best classified in the 'old fact/new guise' category. For, in the open possibility, pre-emergence Mary knew everything there is to know about MP_R, and thus there is a clear and relevant sense in which this guise is not new. At the same

[35] Torin Alter, 'Mary's New Perspective', *Australasian Journal of Philosophy, 73* (1995), 582–4.

[36] One might object that this move generates a vicious regress; for a discussion of this objection see my 'Bats, Brain Scientists, and the Limitations of Introspection', 325–7.

time, prior to emerging from the room, Mary had never represented a phenomenal property by means of this mode of presentation, and thus her deployment of MP_R is new.

5. Phenomenal Concepts and Conceptual Analysis

Thus it seems that until the open possibility of the specified kind of qualitative inaccuracy has been closed off, it can't be claimed that the knowledge argument is sound. To this one might reply that conceptual analysis of our phenomenal concepts reveals that they apply correctly to properties whose qualitative nature is accurately represented by introspection, and that it is ruled out conceptually that they correctly apply to properties whose qualitative nature is not accurately represented in this way. Chalmers suggests an idea of this sort when he specifies that the referent of a pure phenomenal concept is present inside the concept's sense, and its content is constituted by the referent.[37] Some of his physicalist opponents concur. Brian Loar, for example, argues that phenomenal concepts *express* the very properties they pick out. In his framework, a concept expresses its reference-fixer, and thus what he is contending is that reference-fixers of phenomenal concepts are just the properties they pick out. Moreover, he claims that:

> Phenomenal concepts pick out certain properties directly. They do not pick out those properties via a contingent mode of presentation, in the manner say of visual recognitional concepts, which connect one to some external kind by way of a visual experience. It could then seem, I suppose, that phenomenal concepts conceive their referents *as they are in themselves*.

Loar is plausibly interpreted as endorsing the claim that phenomenal concepts accurately represent the qualitative nature of the properties to which they apply.

To evaluate this objection, we first need to be clear about what it is that conceptual analysis reveals. An attractive proposal that derives from

[37] Chalmers, 'The Content and Epistemology of Phenomenal Belief', 13–4. Brian Loar, 'David Chalmers's *The Conscious Mind*', *Philosophy and Phenomenological Research*, 59 (1999), 471; cf. id., 'Phenomenal States', in Ned Block, Owen Flanagan, and Guven Güzeldere (eds.), *The Nature of Consciousness: Philosophical Debates* (Cambridge, Mass.: MIT Press, 1997).

Hilary Putnam is that the structure of certain concepts is a conjunction of conditionals. The antecedents of the conditionals specify coherent scenarios considered as actual—that is, considered as if they were the way things actually turned out, and the consequents indicate what the concept in question then correctly applies to. Which conjunct is *operative* depends on the way the world actually is, since which conjunct is operative is a matter of which conjunct's antecedent is true.[38] This structure is discerned by reflection on possible cases—(Jackson makes an impressive case that the sort of reflection on possible cases that we see in Putnam's work might be thought of as conceptual analysis).[39]

For example, given that all of our samples of the watery stuff in our environment are constituted of H_2O, and this chemical property explains the properties we associate with water, our concept 'water' correctly applies (just) to H_2O, and water = H_2O. But suppose that it had turned out that this watery stuff, like our samples of jade, had two distinct kinds of composition, each at least fairly common. Then claiming that 'water' correctly applies only to H_2O would have been implausible, and, like jade, it would have turned out that water was a disjunctive kind.[40] Or, further, imagine that it instead turned out that this watery stuff had many distinct constitutions with no salient similarities among their intrinsic features, while each sample nevertheless exemplified a well-behaved functional characterization. Then, like 'catalyst' and 'enzyme', we might have correctly counted water as a functional kind. Or, suppose it turned out that Berkeley's view of the universe was correct, and that water was composed just of sensations directly produced in our minds by God. Then we might have classified water as an appearance kind, so that 'water' applied correctly to anything that appeared in one particular way under certain conditions, and in a different set of particular ways under other conditions. If it seems strange

[38] Hilary Putnam, 'The Meaning of "Meaning"', in his *Philosophical Papers*, vol. ii (Cambridge: Cambridge University Press, 1975), 240–1. This idea has been endorsed and developed by George Bealer, 'Modal Epistemology and the Rationalist Renaissance', in Tamar Szabó Gendler and John Hawthorne (eds.), *Conceivability and Possibility* (Oxford: Oxford University Press, 2002), 109; Ned Block and Robert Stalnaker, 'Conceptual Analysis, Dualism, and the Explanatory Gap', *Philosophical Review*, 108 (1999), 36; and Jackson and Chalmers would not dissent from this line of thought, Chalmers and Frank Jackson, 'Conceptual Analysis and Reductive Explanation', *Philosophical Review*, 110 (2001), 322, 340–1.

[39] Jackson, *From Metaphysics to Ethics* (Oxford: Oxford University Press, 1998), 28–86.

[40] The 'would have turned out that S, had it turned out that W' locution derives from Stephen Yablo, 'Shoulda, Woulda, Coulda', in *Conceivability and Possibility*, 454.

that this last scenario has a place in the conceptual analysis of 'water,' it is important to keep in mind that Berkeley's idealist world is not ruled out a priori by conceptual analysis, and that the complete conceptual analysis of 'water' must specify its correct application conditions in any such world, or coherent scenario, considered as actual.[41]

On this proposal, conceptual analysis reveals that our concept 'water' has a structure something like:

> If a scenario is actual in which the watery stuff in the environment has a unique sort of composition, then the concept 'water' correctly applies to a unique compositional stuff;[42]
> and
> if a scenario is actual in which the watery stuff has a small number of sorts of composition, then the concept 'water' correctly applies to a disjunctive compositional stuff;
> and
> if a scenario is actual in which the watery stuff has many sorts of composition, and in which there are no salient similarities among the intrinsic properties of these compositions, while each sample of the watery stuff exemplifies a well-behaved functional characterization, then the concept 'water' correctly applies to a functional kind,
> and
> if a scenario is actual in which each instance of the watery stuff is a collection of sensations produced directly in minds by God, then the concept 'water' correctly applies to an appearance kind . . .

Let us call conjunctions of conditionals of this variety *Putnam-conjunctions*. For certain concepts, the plausibility of this picture serves as a corrective to the idea that conceptual analysis alone can determine, in effect, that a specific conditional is operative. For example, it is sometimes assumed that conceptual analysis alone shows that 'water' refers to a unique compositional stuff. But this would then not be so—in this case, conceptual analysis would reveal only a conjunction of conditionals. Which of the conjuncts is

[41] David Braddon-Mitchell makes a related point in 'Qualia and Analytical Conditionals', *Journal of Philosophy*, 100 (2003), 115. Braddon-Mitchell also suggests that the analysis of some concepts might be a conjunction of conditionals.

[42] The conditionals might also be formulated non-metacognitively, for example: if a scenario is actual in which the watery stuff in the environment has a unique sort of composition, then water is a unique compositional stuff.

operative would then be settled by the actual world, and we would know which conjunct is operative only by our investigation of the actual world. This model allows a concept to remain the same through changes in our scientific theories about what it correctly applies to, or (more salient for present purposes) through a more rudimentary change from a situation in which we rely only on the manifest image for its conditions of correct application to one in which we are informed by a scientific theory. For the model allows that the concept remains the same through such changes, while what is considered to be the operative Putnam-conjunct varies.

Returning to our secondary quality analogy, when examining the nature of color concepts, one might claim that conceptual analysis reveals that

C1. Redness is the property of objects that is the normal cause of their looking red (where 'the normal cause of their looking red' functions merely as a reference-fixer).

But what if it turns out there are many different sorts of causes of looking red, and there are no salient similarities among the intrinsic properties of these causes? C1 might then predict that then there is no such property as redness, or at least that redness is not instantiated—if wildly disjunctive properties are excluded, for example. However, the same might then need to be said about any proposed response-dependent property whose categorical basis was wildly disjunctive, which would be implausible. Then it might turn out that:

C2. Redness comprises whatever properties cause (or could cause) instances of looking red.

Or suppose that because Berkeley's theory turned out to be true, God was the normal cause of objects' looking red. Would we then say that God is red? More likely, redness would then be an appearance property. So then, while conceptual analysis of color concepts might initially seem to reveal something like C1, a more thorough analysis would yield a complex Putnam-conjunction.

There is a moral here for the analysis of phenomenal concepts. Consider the claim that conceptual analysis alone reveals that phenomenal concepts refer to properties that resemble our introspective representations of them, so that

P3. Phenomenal redness is the property that resembles the introspective representation of it.

But by analogy, consider an Aristotelian who holds that conceptual analysis reveals that

C3. Redness is the property of objects that resembles sensations of red.

Suppose he is confronted with a convincing scientific demonstration that physical objects have no such properties. He might conclude that redness is not instantiated in the physical world—that the concept 'red' does not correctly apply to anything in the physical world, as Galileo did.[43] But almost everyone today believes that a response of this sort is mistaken, and that (what is actually) a different conjunct of our concept of red, such as the one that reflects C1, would be operative. Similarly, investigation might lead us to think that the operative conjunct of a phenomenal concept is not the one that has it applying correctly to a property whose qualitive nature is accurately represented introspectively. Suppose it turns out that there are no instantiated properties accurately represented by introspective phenomenal representations. One might conclude that phenomenal concepts fail to apply to any instantiated properties. However, as in the case of color concepts, a radical conclusion of this sort is not clearly forced on us. Rather, it might well be that there are alternatives reflected in other conjuncts in the analysis of phenomenal concepts.

It might be objected that while it is possible to devise phenomenal concepts whose analysis is complex in this way, still our ordinary phenomenal concepts are simple in the sense of being non-conjunctive, and have something like P3 as an analysis, so whether there are phenomenal properties on the ordinary understanding depends on whether there are properties that fit something like P3. One might envision Aristotelians about color having made an analogous claim: 'One might devise color concepts whose analysis is a complex conjunction of conditionals, but ordinary color concepts are simple, and are to be analyzed on the order of C3.' But, as history has shown, the initial attractiveness of C3, and its resilience, are insufficient to show that it provides the complete and correct analysis of our concept of red. The case of phenomenal redness is, I suggest, parallel. The initial attractiveness of something like P3, and its resilience, are insufficient to show that it is the complete and correct analysis of our

[43] Galileo Galilei, *The Assayer*, in *Discoveries and Opinions of Galileo*, trans. Stillman Drake (New York: Doubleday Anchor, 1957), 274–7.

concept of phenomenal redness. Rather, it is an open possibility that the analysis of this concept reveals a complex Putnam-conjunction, and that the operative conjunct renders true a different sort of characterization, such as:

P1. Phenomenal redness is the property that is the normal cause of introspective representations of phenomenal redness (where 'the normal cause of introspective representations of phenomenal redness' functions merely as a reference-fixer).

or

P2. Phenomenal redness comprises whatever properties cause (or could cause) instances of the introspective representation of phenomenal redness.

Even if typical current theories of phenomenal concepts attribute qualitative accuracy to them, dispensing with those theories might well not be ruled out by the nature of the concepts themselves, and may be welcome. The Aristotelians maintained that our concept of temperature is a sensory concept, and that it generally represents the qualitative nature of temperature accurately. By contrast, Locke argued that our temperature concept is a secondary quality idea—a type of response-dependent concept—and that it or its sensory content represents the qualitative nature of temperature inaccurately.[44] A Kripkean understanding has it that our concept of temperature, like our concept of water, is not a secondary quality concept, but that it is rather a natural kind concept, and again—at least given Lockean intuitions—that our tactile representations of temperature are qualitatively inaccurate. Plausibly, this evolution of theory about our concept of temperature amounts to progress—perhaps we have a better understanding of which conjunct of the Putnam conjunction for 'temperature' is actually operative than the Aristotelians did. Similarly, even though it may currently be compelling to theorize that our (pure) phenomenal concepts are accurate in their representation of the qualitative nature of phenomenal properties, we may be led to conclude otherwise. It is an open possibility that a causal account of phenomenal representation is true, and this gives rise to the open possibility that phenomenal concepts are represent phenomenal properties as having qualitative natures they do not in fact have. Then we may find that for each type of introspective phenomenal representation there is a single type of physical property that is its normal underlying

[44] Locke, *An Essay Concerning Human Understanding*, II. viii, 15.

cause, whereupon we might conclude that phenomenal concepts correctly apply to such underlying physical causes. Or else we may find that there are many very different sorts of properties that can cause instances of a single type of phenomenal property, and this might have us come to believe that phenomenal concepts correctly apply to any such properties. The conceptual analysis of 'phenomenal redness' might well allow for such alternatives, despite what we may have thought. If one wants to deny this, and claim instead that by conceptual analysis it can be shown that just

P3. Phenomenal redness is the property that resembles the introspective representation of it

is generally representative of the analysis of the relevant sort of phenomenal concepts, it seems to me that one would need to develop more thoroughly a theory of such concepts (such as the self-presenting or constitution views) that would indicate how it might be that this claim is clearly true.

6. Edenic and Ordinary Phenomenal Content

The analogy with secondary quality representation can be developed further to strengthen the challenge from the qualitative inaccuracy hypothesis to the intuition that Mary learns something new upon seeing the tomato. Consider Chalmers's view of the content of phenomenal color representation. He argues that the account of such content that is most adequate to the phenomenology of color perception is primitivism (of which the Aristotelian position is a variety):

> The view of content that most directly mirrors the phenomenology of color experience is primitivism. Phenomenologically, it seems to us as if visual experience presents simple intrinsic qualities of objects in the world, spread out over the surface of the object. When I have a phenomenally red experience of an object, the object seems to be simply, primitively, *red*. The apparent redness does not seem to be a microphysical property, or a mental property, or a disposition, or an unspecified property that plays an appropriate causal role. Rather it seems to be a simple qualitative property, with a distinctive sensuous nature. We might call this property perfect redness: the sort of property that may have been instantiated in Eden.[45]

[45] Chalmers, 'Perception and the Fall from Eden', in Tamar Szabó Gendler and John Hawthorne (eds.), *Perceptual Experience* (Oxford: Oxford University Press, 2006), 66.

Rather than characterizing primitive properties by way of resemblance to sensations, as Locke does, Chalmers opts for characterizing them as properties that are as they seem to sensory experience. To sensory experience these properties appear simple in the sense of not having an internal causal or dispositional structure, and in the sense of not being composed, for example, of microphysical particles. They also appear to be non-mental properties. In addition, our experience of them is not as unspecified properties—we might say that we experience them as having a specific and determinate nature.[46] The content of a phenomenal color representation associated with these primitive properties Chalmers calls its *Edenic content*.[47]

But Chalmers thinks that science and philosophical reflection provide us with good reasons to believe that there is no instantiated property to which this Edenic content correctly applies—there are no instantiated primitive color properties. However, this does not mean that there are no colors. For there is a veridical content of phenomenal color representation that well-enough *matches* its perfect content—which Chalmers calls its *ordinary content*. Edenic content functions as a kind of regulative ideal in determining the ordinary content of our color experiences—it is the standard that matching ordinary content must most closely approximate—but its being merely a regulative ideal allows for matching ordinary content that is veridical.[48] Note that this account seems to commit Chalmers to qualitative inaccuracy in visual color perception. Such perception represents physical objects as primitively colored, but they are not primitively colored, while at the same time they are colored. Visual color perception sorts colors quite correctly, but represents something else about them inaccurately, and the only candidate for what is inaccurately represented would seem to be the color's qualitative nature.

But notice that a story parallel to Chalmers's account of the content of color representation can be given for our introspective representations of phenomenal properties. When I have an introspective representation of phenomenal redness, what I apprehend seems to be simply, primitively,

[46] In the Garden of Eden, Chalmers specifies, 'we had unmediated contact with the world. We were directly acquainted with objects in the world and with their properties. Objects were simply presented to us without causal mediation, and properties were revealed to us in their true intrinsic glory'; 'Perception and the Fall from Eden', 48. Chalmers is specifying an ideal here; he does not deny that primitivism about visual color representation can accommodate a causal theory of such representation, as in the Aristotelian view.

[47] Ibid., 69–71.

[48] Ibid., 69–84.

phenomenally red. Phenomenal redness seems to be a simple qualitative property, with a distinctive sensuous nature. We might call this property primitive phenomenal redness, and the content of introspective phenomenal redness associated with this property its Edenic content. But it may be that these representations are also qualitatively inaccurate in the sense that their Edenic content correctly applies to no instantiated properties. Still, there might be an ordinary content of these representations that matches their Edenic content closely enough, with the consequence that there are instantiated phenomenal properties to which this matching content correctly applies. These properties might be physical properties, such as the physical property that is the normal cause of introspective representations of phenomenal redness.

Consider two possible proposals for the ordinary content of representations of phenomenal redness (derived from P1 and P2 above):

OCi: an ordinary content that correctly applies to the property that is the normal cause of introspective representations of phenomenal redness (where 'the normal cause of introspective representations of phenomenal redness' functions merely as a reference-fixer).

OCii: an ordinary content that correctly applies to whatever properties cause (or could cause) instances of the introspective representation of phenomenal redness.

On OCi, it is nomologically possible for a state to be introspected as phenomenal redness while phenomenal redness is not then instantiated, since a property that on some occasion causes the introspective representation of phenomenal redness might not be the property that is its normal cause. Accordingly, a characteristic of OCii that might argue in favor of its being the closest match to the regulative ideal in that it would make it impossible for a state to be introspected as phenomenal redness while phenomenal redness is not instantiated—so that if a state seems conscious in the what-it-is-like-to-see-red way, it will be conscious in this way. As applied to pain, for example, this might count in favor of OCii, since it is at least initially strongly unintuitive for many that a state be introspectively represented as pain and not be pain. On the other hand, it may count in favor of OCi that it would allow our classification of phenomenal properties to cut nature at its causal joints, after the manner of Kripkean natural kind terms or concepts, while OCii is not designed to do so. I favor OCi. As for its unintuitive consequence, I suspect that it might well be that a state be

introspectively represented as pain and not be pain. Consider the reverse of the Novocain example discussed earlier: someone says that he is going to inject a large needle into your mouth, but instead administers drops of cold water. If, introspectively, pain can be mistaken for the sensation of drops of cold water, then it would seem that, introspectively, the sensation of drops of cold water could be mistaken for pain. So then a state might be introspectively represented as being pain, but not be pain.[49]

Given an Edenic content interpretation,

(T) Phenomenal redness has qualitative nature Q,

and supposing (T) is true, Mary would learn something new when she leaves the room and sees the red tomato. But, on the open possibility we are considering, on an Edenic content interpretation, (T) is in fact false. So then Mary would not learn something new. On several ordinary content interpretations (for example, OCi and OCii), it turns out that (T) would be derivable from what pre-emergence Mary knows at least there would then be no less reason to believe that (T) is derivable from this information than there is to think that 'more than half of the earth's surface is covered with water' is so derivable. Then again, Mary would not learn anything new when she sees the tomato, but for a different reason—she already knew (T) when she was in the room. Thus, in our open possibility, for both Edenic and ordinary content interpretations of (T), Mary does not learn anything new when she leaves the room, and the knowledge argument faces a challenge.

7. The Explanatory Gap and Eliminativism

Chalmers contends that several commentators who have attempted to undermine the knowledge argument (and the zombie argument) by the 'old fact/new guise' response have failed to show how it might be that the distinct modes of presentation, physical and phenomenal, might refer to the same thing. This issue is especially pressing for Loar, who holds that a phenomenal concept expresses the phenomenal property to which it refers, while physicalism is true. On Chalmers's reading, Loar in fact maintains 'that

[49] Thanks to Kati Balog for raising the objection that occasioned these thoughts.

phenomenal and physical concepts (i) are cognitively distinct, and (ii) both express the property they refer to'.[50] Chalmers argues that if both (i) and (ii) are accepted, nothing can justify the claim that the phenomenal and physical concepts co-refer. Or perhaps it is actually impossible that (i) and (ii) are both satisfied, for the reason that the phenomenal concept correctly applies to a primitive phenomenal property, while this is not the property the physical concept expresses. However, if introspective phenomenal representations are qualitatively inaccurate, Chalmers's explanatory burden—which is part of the burden of the explanatory gap—can be discharged. For then it need no longer be explained how a qualitative nature that resembles the what-it-is-like-to-sense-red mode of presentation can *correctly* be attributed to a phenomenal property, while that property is at the same time physical, and a description of its real qualitative nature is represented by or derivable from 'P'.

With regard to this explanatory gap, Adams argues that materialism has no adequate response to the demand to explain why particular kinds of phenomenal properties are correlated with particular kinds of physical properties:

> For suppose a materialist claims that [physical property] *R* and the phenomenal appearance of red are one and the same property of brains, identified as *R* on the basis of its place in the physical system, and as the appearance of red on the basis of the way it seems to us when our brains have it. We can still ask why *R* seems to us the way it does, rather than the way *Y* (the physical brain state which 'is' the appearance of yellow) does. This is quite recognizably our original question, and it remains unanswered.[51]

Adams's demand is for a contrastive explanation: why does physical property *R* seem the way it does, and not the way physical property *Y* seems? Supposing that different brain states appear in different ways to introspection, the physicalist needs to explain why any one brain state appears to introspection in one way, and not in some other way—for example, in the way some other brain state appears to introspection. Adams believes that the physicalist has no adequate response to this demand.

But we can reply: it is an open possibility that there is a discrepancy between the real nature of the property *how* R *seems* and the qualitative

[50] Chalmers, 'Materialism and the Metaphysics of Modality', 487–8.
[51] Adams, 'Flavors, Colors, and God', 259.

nature we introspectively represent this property as having. It is then an open possibility that *how* R *seems* is a straightforwardly physical property, call it RS, despite how we introspectively represent it. The same can be said of *how* Y *seems*—it might be a straightforwardly physical property—call it *YS*, despite the qualitative nature we introspect it as having. If this open possibility is actual, then the physicalist can meet the demand for contrastive explanation, which might then be formulated as: why does physical property *R* cause physical property *RS* and not physical property *RY*? We're assuming that Mary, while she is in the room, has mastered purely physical explanations of this sort. So, on the open possibility under consideration, the physicalist can meet Adams's demand for an explanation.[52]

Adams further contends that a materialistic explanation of correlations between physical and phenomenal properties would require a materialism of a radical sort:

> one would have to adopt a very radical materialism indeed, rejecting not only the dualism of substances, but also the dualism of properties, and even the distinction of first- and third-person aspects or ways of identifying the sensible qualities, as well as the notion of a way in which conscious states seem to us when we are in them, as opposed to their place in the physical scheme of things. Thus one would have to eliminate phenomenal qualia, or reduce them in a most extreme way to physical qualities.[53]

However, the materialism suggested by our open possibility can retain the distinction between first- and third-person ways of identifying the sensible qualities, and also the notion of a way in which conscious states seem to us when we are in them. For, despite the discrepancy between the real qualitative nature of phenomenal properties and how they are introspectively represented, there is a first-person, introspective point of

[52] As mentioned in n. 7, Stoljar's version of the knowledge argument specifies that pre-emergence Mary possesses all the phenomenal concepts, while she nevertheless lacks knowledge of how correct applications of phenomenal concepts are correlated with physical states ('Physicalism and Phenomenal Concepts'). The anti-physicalist might then contend that while she is in the room Mary would not be able to produce contrastive explanations of the sort Adams demands, and that for this reason physicalism is false. But now we can see that on the open possibility under investigation Mary *would* be able to produce these explanations. One might press on here by asking: why does *RS* appear to us as it does and not otherwise? Here again we can ascend a level and suggest the open possibility that our introspective representation of the appearance of *RS* (*ARS*) is inaccurate, and that *ARS* is a physical property. Then, by virtue of having mastered all the physical explanations, Mary understands why *RS* causes *ARS*, and not some other relevant alternative physical property.

[53] Adams, 'Flavors, Colors, and God', 259.

view on phenomenal properties, and a way they appear to us when we are in the states that have them. True, the real qualitative nature of those properties would be accessible from the third-person point of view, so the first-person perspective does not provide genuine information about the qualitative nature of those properties that is not accessible from the third-person perspective. But this is not enough to make the materialism in question a radical one, since a claim of this sort would be required for any materialism.

At the same time, denying that a qualitative nature that resembles introspective phenomenal modes of presentation is correctly attributed to phenomenal properties might well not amount to eliminativism for phenomenal properties. One could, in agreement with Galileo, argue for eliminativism about temperature as a property of physical objects, on the grounds that the temperature of physical objects does not resemble our sensory representation of it.[54] An Edenic content of our temperature concept could even be defined that applies only to a property that resembles temperature sensations—and it might then be pointed out that there is no actually instantiated property to which this content correctly applies. But, as history has shown, highly plausible non-eliminativist options for temperature itself remain. On the open possibility that I have been discussing, non-eliminativist options also remain for phenomenal properties. *Something* that many believe to exist would be eliminated—certain features that are accurately represented introspectively. Indeed, one might define a notion of the Edenic content of phenomenal concepts that would correctly apply only to such features, which would then correctly apply to no properties that are actually instantiated. But this is not to say that phenomenal properties would thereby be eliminated, or that our phenomenal concepts fail to apply to anything real, for they might have an ordinary content that does. Even then, the Edenic content might still function as a regulative ideal, on the model for color concepts Chalmers develops.

8. The Big Picture

Against the knowledge argument I have contended that since it is an open possibility that the right account of introspective representation is

[54] Galileo, *The Assayer*, 274–7.

causal, it is also an open possibility that introspective representations of phenomenal properties are in a sense qualitatively inaccurate. Specifically, we introspectively represent phenomenal properties as having a certain qualitative nature, but the attribution of this qualitative nature to these properties might be incorrect. As a result, it may be that the real nature of phenomenal properties is straightforwardly physical, and complete information about it is derivable from what Mary knows before emerging from the room, despite the appearance that she acquires information about the qualitative nature of phenomenal-redness when she leaves the room that was not available to her earlier. Kant might well have endorsed the open possibility of this kind of qualitative inaccuracy.[55] By contrast, Descartes arguably maintains that qualitatively accurate and complete introspective representations of our mental states are generally available to us.[56] Perhaps we should say that both the Cartesian and Kantian options, applied to phenomenal states, are live open possibilities; neither has been ruled out. But as long as the Kantian open possibility remains standing, the knowledge argument against physicalism faces a significant challenge.

[55] Kant, *Critique of Pure Reason*, e.g., Bxxiv–xxvii.

[56] Descartes, *Meditations on First Philosophy*, *The Philosophical Writings of Descartes*, ii 16–23.

5

Kant on Apriority and the Spontaneity of Cognition

HOUSTON SMIT

In his superb *Leibniz: Determinist, Theist, Idealist*, Robert Adams observes that Leibniz usually uses 'a priori' in its original sense, one with roots in Scholastic Aristotelianism.[1] On this notion, to prove something a priori is to prove it from its causes, where 'cause' is understood very broadly to mean 'explanatory ground'. To prove something a posteriori, in contrast, is to prove it from its effects, or consequences. Let's refer to this now-archaic notion of the a priori as the 'from-grounds' notion. The from-grounds notion of the a priori clearly has a metaphysical orientation absent in the epistemic notion familiar to us, on which apriority concerns simply the justificatory independence of knowledge from experience. It also, intriguingly, differs from the non-empirical notion of the a priori in that it offers a positive characterization of what it is to know a priori: to know a priori is to know through an a priori proof, and so from the grounds that explain what is known.

In attributing the from-grounds notion of the a priori to Leibniz, Adams suggests that Leibniz plays a crucial role in the transformation of the sense of 'a priori' to the epistemic one familiar to us.[2] Leibniz held that experience provides a basis for knowledge only as an effect, or consequence, so that to know something from its cause, or explanatory ground, is to know it independently of experience. According to Adams, this explains why Leibniz appears at points to use the term 'a priori' to mean 'non-empirical': these are contexts in which the idea of independence from experience is foremost

[1] See Robert M. Adams, *Leibniz: Determinist, Theist, Idealist* (New York: Oxford University Press, 1994), 109–10.

[2] Ibid., 110.

in his mind. Those who shared Leibniz's epistemological views, such as Alexander Gottlieb Baumgarten, then found it natural to *define* the a priori as the non-empirical. Adams also suggests that the from-grounds notion of the a priori enjoyed widespread currency in the seventeenth century, and that it still enjoyed currency among German philosophers, including Christian August Crucius, into the middle of the eighteenth century.[3]

All of these suggestions are, I think, extremely important ones. In this paper, I want to take them in a direction that may strike the reader as surprising. Although—as Adams, and others, hold—Kant played a decisive role in giving the non-empirical notion of the a priori the dominance that it enjoys to this day, I believe that the from-grounds notion of the a priori is none the less also operative in Kant's critical work. In this paper, I will argue that it is, in particular, operative in the *Critique of Pure Reason* in much the way that Adams finds it to be present in Leibniz's writings—as the primary, and governing, notion that entails, among others, the idea of independence from experience.

Seeing that Kant works with the from-grounds notion of the a priori is important, because it brings into clear relief how the central claims and arguments of the first critique have a crucial metaphysical, as well as an epistemological, dimension. This work aims to supply a determinate account of the only order of ontological grounds that humans, given the nature of their cognitive capacity, can be conscious of in having a priori theoretical cognition. It aims to revolutionize metaphysics by proving that objects of our experience depend for their possibility on an ontologically prior order or structure inherent in the human mind. This dependence parallels, in crucial respects, the dependence that, according to Leibniz, the possibility of created existent things have on God's existence, via necessary truths in God's mind:

> And lest you should think that it is unnecessary to have recourse to this [the Supreme and Universal] Mind, it should be borne in mind that these necessary truths contain the determining reason and regulating principle of existent things—the laws of the universe, in short. Therefore, since these necessary truths are prior to the existence of contingent beings, they must be grounded in the existence of a necessary being.[4]

[3] Ibid.

[4] Leibniz, *New Essays on Human Understanding*, trans. Peter Remnant and Jonathan Bennett (New York: Cambridge University Press, 1996), 447. Hereafter, citations from this work will be given

To appreciate how the critical turn identifies the human mind, instead of the divine, as what in this way contains the a priori grounds of the possibility of objects of experience in general is to appreciate how the *Critique of Pure Reason* really is, in conception and execution, a work of metaphysics, and not just of epistemology.[5]

The full case I ultimately want to build for reading Kant as operating with the from-grounds notion of the a priori, then, turns on elaborating how ascribing this notion of the a priori to him sheds light on the central claims and arguments of his critical philosophy. Providing a full case for this reading is, however, far too ambitious a project for the present piece. What follows is a first installment on this project. It explains how ascribing the from-grounds notion of the a priori to Kant clarifies the array of different characterizations of the a priori in the *Critique of Pure Reason*. It also explains how Kant puts this notion to work in posing the fundamental problem addressed in the first critique, the problem regarding the a priori objective validity of the categories that he sets out to solve in the Transcendental Deduction of the Categories.

I will focus much of my attention on what is perhaps the most difficult, and important, variation in Kant's use of the term 'a priori' in the first critique, one that many interpreters have regarded as an outright ambiguity: 'a priori' seems sometimes to refer to a cognition's justification, and at others, however, to its genesis.[6] Recognizing that Kant works with the from-grounds notion of the a priori will clarify how the relevant genesis

parenthetically, using 'NE' as an abbreviation, e.g., (NE 447). The following abbreviations will be given for Leibniz's works: C for *Opuscules et fragments inédits*, ed. L. Couturat (Paris: Félix Alcan, 1903); DM for the *Discourse on Metaphysics*, trans. Ariew and Garber in their *Philosophical Essays* (Indianapolis, Ind.: Hackett, 1989), which will in turn be abbreviated by AG; G for *Philosophische Schriften von G. W. Leibniz*, ed. C. I. Gerhardt (Berlin, 1875–90); L for *Philosophical Letters and Papers*, ed. L. E. Loemker, 2nd edn. (Dordrecht: Reidel, 1969).

[5] I should note that other scholars, such as Henry Allison, have suggested that the critical turn gives the human mind a role traditionally accorded to the divine. See his *Kant's Transcendental Idealism* (New Haven, Conn.: Yale University Press, 1983), 28–9. However, the reading I am proposing is incompatible with Allison's influential reading of Kant's transcendental conditions on the possibility of experience as mere 'epistemic conditions' (ibid.). These conditions themselves constitute a priori cognition of things, and ground the possibility of all our cognition of things, only as a priori grounds of the possibility of these things in the from-grounds sense of 'a priori' and so as grounds of the possibility of these things themselves, and not merely as grounds of the possibility of our knowledge of these things.

[6] Patricia Kitcher, for example, distinguishes different senses of 'a priori' in Kant, including a justificatory and a genetic one. See her *Kant's Transcendental Psychology* (New York: Oxford University Press, 1992), 15–17.

and justification are different aspects of a single positive conception of a priori cognition. Indeed, a central contention of Kant's critical philosophy is that our a priori cognition—cognition which registers an ontological ground as such, and in which that ontological ground thereby comes also to serve as a cognitive ground—is cognition that 'our own cognitive capacity (merely prompted by sensible impressions) provides out of itself [*aus sich selbt hergibt*]'.[7] In examining the sense and motivation of this genetic characterization of a priori cognition—which, as will emerge, is itself an a priori cognition, in the from-grounds sense, of the possibility of our a priori cognition—it will prove helpful to attend to Kant's characterization of the understanding as 'the capacity for bringing representations forth itself, or the spontaneity of cognition'.[8]

1. The From-grounds Notion of the A Priori

I want to begin by explicating the original, from-grounds notion of the a priori that Kant inherits. As Adams points out, Arnauld and Nicole report the widespread currency this notion enjoyed in the seventeenth century when they claim, in their influential *Port Royal Logic*, that our minds are 'capable of finding and understanding the truth, either by proving effects by the causes, which is called an a priori proof, or, on the contrary, by demonstrating causes by the effects, which is called an a posteriori proof'.[9] In explicating this notion, however, I will be attending mainly to Leibniz's treatments of the a priori, because of his influence, direct and indirect, on Kant. Leibniz expresses this notion of the a priori explicitly when he equates knowledge a priori—which presumably includes knowledge through a priori proof—with knowledge through causes.[10] We will also

[7] B1. The standard German edition of Kant's works is *Kant's Gesammelte Schriften*, edited by the Royal Prussian (later German) Academy of Sciences (Berlin: Georg Reimer, later Walter de Gruyter and Co., 1900–). Citations from the *Critique of Pure Reason* will be given using the pagination of this edition, using 'A' to specify the 1781 version and 'B' to specify that of 1787: so, for example, (B1) refers to the first page of the Academy edition of the 1787 version. All other citations from Kant's works will specify the volume and page numbers, separated by colon, of the Academy edition. The translations from the first critique are, with minor variations, drawn from P. Guyer and A. Wood's translation (Cambridge: Cambridge University Press, 1998); and many of the translations from Kant's lectures on metaphysics are drawn from K. Ameriks and S. Naragon's translation (New York: Cambridge University Press, 1997).

[8] A51/B75.

[9] IV, 1; as cited by Adam, *Leibniz*, 109.

[10] C 272; *Leibniz*, 109.

see that the pre-critical Kant employs the from-grounds notion of the a priori in the *Nova Dilucidatio*, albeit without employing the term 'a priori'.

In explicating the from-grounds notion of the a priori, I will be attending specifically to a priori proof and knowledge. Doing so will serve to bring out the connection this notion of the a priori bears to reason and necessary truth, a connection implicit in the general rationalist conception of scientific knowledge shared by Leibniz, Wolff, and Kant. The reader should keep in mind, however, that Kant does not speak only of proving (*beweisen*) a priori (as for instance, at A774/B802) and knowing (*wissen*) a priori (as for instance, at B2). He also, and more commonly, speaks of cognizing (*erkennen*) a priori, or of a priori cognition (*Erkenntnis*). What is more, he predicates apriority to a wide array of different subject-matter, in addition to knowledge and cognition—consciousness, rules, intuitions, concepts, judgments, truths, principles, construction, exhibition, synthesis, sensibility, possibility, grounds, certainty, and the origin or genesis of representations. For our purposes, the crucial thing to clarify will prove ultimately to be the apriority of cognition, and specifically how Kant finds the a priori grounds of this apriority in a cognition's deriving from an exercise of the spontaneity of cognition.

The point most crucial to understanding the from-grounds notion of the a priori is one that Adams helpfully stresses: what distinguishes an a priori proof, in the from-grounds sense, from an a posteriori one is that only the former *explains* what it proves.[11] Leibniz remarks that 'Proof a priori or *Apodeixis* is explanation of the truth',[12] implying that it is only in proving a proposition from the cause, or ground, that makes it true that one explains *why* it is true, as against merely establishing *that* it is true. That, for Leibniz, what distinguishes a priori proofs is their explaining what they prove is also evident when he classifies indirect proofs—proofs that proceed by deriving a contradiction from the negation of what is to be proved—as a posteriori:[13] Leibniz evidently counts indirect proofs as a posteriori, on the grounds that they fail to explain what they prove.[14]

[11] *Leibniz*, 77. [12] C 408. [13] C 154.

[14] This paragraph closely follows Robert Adams's discussion on p. 109 of *Leibniz*, where he cites these texts in support of his contention that Leibniz employs the notion of the a priori that is to be found in the *Port Royal Logic*. It is worth noting that the *Port Royal Logic* counts Euclid's proofs by reduction to absurdity as defective, on the grounds that they do not explain the truth of what they prove, by appeal to the principles of the relevant thing (IV, 9).

Notice that in classifying indirect proofs as a posteriori, Leibniz counts the contradiction from which one reasons back to the truth of what is proved, as an effect of this truth. He thus understands 'effect' very broadly, indeed: the effect from which an a posteriori proof proceeds need not be, in our restrictive sense, a causal effect. Correspondingly, the term 'cause' in these passages needs to be understood as 'ground' in the broad sense of an 'explanatory factor', and 'effect' as 'consequence' in a correspondingly broad sense. To reflect this point—and in anticipation of Kant's insistence that only a particular species of ontological ground is to be called a 'cause' (*causa, Ursache*)—I have dubbed the original sense of 'a priori' the 'from-grounds', and not the 'from-causes', sense.

Knowing the truth of a proposition a priori, then, requires not just establishing *that* the proposition is true, by appeal to grounds that suffice to show that it is true. It requires, rather, knowing *why* that proposition is true, by grasping or reasoning to that proposition from the grounds that make it true. There are two points here that are worth bringing out explicitly. The first is that a priori proof appeals, not just to any ground that provides evidence for what is to be proved, but to grounds that are prior, in the order of being, to what is proved. Knowing a priori requires appeal, we might say, not simply to any ground of knowledge, or epistemic ground, but to the ground that makes what is known the case, an ontological ground. The second point is that a priori knowledge requires rational perception of the necessity with which an ontological ground determines, and thereby explains, what is known. Indeed, it is in virtue of such a perception that an ontological ground comes to constitute a ground of one's a priori knowledge.

A paradigm case of a priori knowledge, on this understanding of the a priori, is the Euclidian geometer's knowledge of a given figure's necessary properties. For example, the Euclidian geometer knows a priori that the sum of the internal angles of a triangle is equal to the sum of two right angles. She does so when she perceives, in and through the execution of certain acts of geometric construction, how the essential properties of a triangle, specified in its definition, together with certain postulates, make it the case that this truth obtains, and obtains necessarily. But someone might come to know—at least in a suitably permissive sense of 'know'—the truth of the same theorem a posteriori, by carefully drawing various triangles and measuring the sum of their internal angles using a protractor. Our protractor-wielder postulates that the essence of a triangle is the ontological

ground that determines what he observes, and so in establishing the theorem proceeds from the consequence to the ground. What he lacks, and the geometer enjoys, is just the rational perception of necessity with which the essence of a triangle determines the truth of this theorem—perception that constitutes knowledge of why the theorem obtains universally and necessarily.

It will prove useful to see how Leibniz draws this connection between the a priori and reason in distinguishing 'truths a priori, or of Reason' from 'truths a posteriori, or of fact'.[15] This contrast between truths a priori and truths a posteriori is one between truths considered in so far as they are properly, or ideally, known, either by us or more generally by finite minds.[16] Truths that we properly, or ideally, know a priori, such as those of geometry, are truths of reason, because we know them properly through reason—a discursive faculty for appreciating real grounds as such and, thereby, *why* certain propositions are true. Leibniz calls a posteriori truths 'truths of fact', in contrast, to indicate that we cannot see why they are true: as far as we can see, they are truths of mere fact.

In Leibniz's account, the knowledge of a finite mind has two fundamental sources: 'the senses and reflection'.[17] The senses yield perception, and reflection yields intelligence: 'There are two sorts of knowledge: that of facts, which is called perception, and that of reasons, which is called intelligence. Perception is of singular things, intelligence has for its objects universals or eternal truths.'[18] Perception is of singular things in virtue of representing the infinite number of predicates that, in Leibniz's view,

[15] NE 434.

[16] This is a point made by Tyler Burge in his important 'Frege on Apriority', in Paul Boghossian and Christopher Peacocke (eds.), *New Essays on the A Priori* (Oxford: Oxford University Press, 2000), 18–19. Burge contends that, for Leibniz and Frege, what are a priori or a posteriori are truths considered in respect of their associated 'ideal or canonical justification' (ibid.). Moreover, he points out that Kant's treatment of the a priori differs from Leibniz's and Frege's in this respect: for Kant, the a priori and a posteriori are primarily to be predicated, not of truths, but of 'cognition and the employment of representations' (ibid.). The central thesis of the present paper, in effect, specifies what this shift Burge points out consists in, and provides an explanation for it: for Kant, cognition and the employment of representations are that of which the a priori and a posteriori are to be predicated, because the only order of ontological dependence that we, and more generally any subject of discursive understanding, can register in cognition, be the cognition a priori or a posteriori, is not an ontological order that, considered transcendentally, obtains independently of how we represent and cognize things. Notice that, in the present reading, Kant's understanding of the distinction between a priori and a posteriori cognition retains the ontological orientation it has in Leibniz; it is simply that this ontological orientation regards an order that is, when considered transcendentally, merely ideal.

[17] NE 53.

[18] G 583.

individuates a singular thing. A finite mind can represent an infinite number of predicates only confusedly, and so its perception, through sense, of singular things must be confused. Intelligence, in contrast, is by definition not only clear, but distinct.[19] For this reason, the intelligence of a finite mind has only universals or eternal truths as its object. And since a finite mind's knowledge of a priori truths is a species of intelligence, it follows that it cannot have a priori knowledge of singular things as such. In particular, a finite mind's intelligence has a priori truths as its object, in so far as it has essences, the possibilities of things,[20] and eternal truths, the necessary relations among essences,[21] as its objects. For a finite mind knows a priori truths in analyzing these essences, and drawing on the eternal truths, to grasp how the possibilities of things are the ontological grounds of these truths.[22]

To be sure, according to Leibniz, there are a posteriori truths that we can know to be universally true, in a less demanding sense of 'know'. These are truths that concern contingent beings in so far as they share general natures evident to us through experience. He thus remarks that a posteriori truths can 'also become universal, in a way, but that is by induction, or observation, so that what we have is only a multitude of similar facts, such as the observation that all quicksilver is evaporated by the action of fire'.[23] And he goes on to remark,

> This is not perfect universality, since we cannot see its necessity. General propositions of reason are necessary, although reason also yields propositions which are not absolutely general, and are only likely—for instance, when we assume that an idea is possible until a more accurate inquiry reveals that it is not.[24]

As this suggests, what is crucial to Leibniz's distinction between a priori and a posteriori truths is that the former are (at least ideally), whereas the latter are not, knowable through reason by a rational perception of their necessity.[25] It is this rational perception of essences as they contain the

[19] NE 173. [20] NE 293. [21] L 488.

[22] Donald Rutherford cites these texts in the course of a valuable discussion of Leibniz's conception of metaphysics and its methodology, one to which my discussion is indebted. See ch. 4 of his *Leibniz and the Rational Order of Nature* (New York: Cambridge University Press, 1995).

[23] NE 446. [24] Ibid.

[25] I should stress that Leibniz does not claim that all a priori knowable truths are necessary truths: in his view, an infinite mind can establish contingent truths, and thereby know them as certain, by a priori proofs (DM § 13). All he claims is that truths that can be proved through reason, and that thus are in principle knowable a priori by finite minds, are necessary truths. Any a priori proof of a truth

ontological grounds of a necessary and eternal truth, and so in respect of how they explain that truth's obtaining, and obtaining necessarily, that constitutes our certain and perfect knowledge of that truth's universality. For Leibniz, then, a priori truths are truths about possibility and necessity that a finite mind can, at least ideally, know in this rational perception of essences, from the grounds that make them true. It is a consequence of their being truths that can be so known, that a priori truths are truths that a finite mind can with certainty know to be necessarily and so universally true.[26]

For our purposes, it will also be important to see how Leibniz's view that essences are the objects of a finite mind's a priori knowledge restricts the scope of demonstrative sciences to possibilities and necessities: The real existence of beings which are not necessary is a matter of fact or of history, while the knowledge of possibilities and necessities (the necessary being that the opposite of which is not possible) is what makes up the demonstrative sciences.[27] Purely demonstrative sciences, such as pure mathematics, are concerned with possibilities and necessities because they are concerned solely with a priori truths. However, Leibniz holds that there are also sciences, such as astronomy, that concern 'mixed propositions' that derive from a priori and a posteriori truths. The necessities that these sciences deal with are, presumably, only hypothetical, and it is only in so far as these sciences apply eternal truths to contingent beings in proofs that are only impurely a priori that they count as demonstrative. Astronomy, for instance,

reveals its real ground—on Leibniz's view, the predicate's containment in the subject. In the case of a necessary truth, this containment can be revealed in a finite analysis that resolves that proposition into manifest truths of identity. But in the case of a contingent truth, the predicate is contained in the subject only in virtue of the infinitely complex content of the subject's individual concept. The a priori proof of a contingent truth, then, requires intuition or deduction (in Descartes' senses) of this infinitely complex content, something only an infinite mind is capable of. Such a proof cannot, in particular, be supplied through reason, since its analysis of this content is a discursive process that could never reveal its full, infinite, content. For an illuminating discussion of Leibniz's infinite analysis account of contingency, see John Carriero, 'Leibniz on Infinite Resolution and Intra-Mundane Contingency', in two parts: *Studia Leibnitiana*, 25 (1993), 1–26 and ibid., 27 (1995), 1–30.

[26] Notice the implication of NE 446 that a posteriori truths that concern contingent things in so far as they share natures that are evident to us in experience are necessary. Leibniz's thought, I take it, is that since such a truth concerns a general kind, its subject is a concept that consists of a finite number of predicates, so that the containment of its predicate in these concepts could be revealed in analysis, and so demonstrated. The problem is that we can derive concepts of these general kinds only from sense—we cannot derive them solely from reflection—and sense yields to our intelligence only the conjunction of these predicates in a concept. It does not, in particular, yield to intelligence the ontological ground (ultimately the essence) that constitutes the possibility of that concept and that renders an a posteriori truth of the kind in question necessary. And this is just to say that we cannot see the necessity of such an a posteriori truth.

[27] NE 301.

provides such impurely a priori proofs when it applies mathematical theorems to observations of the stars to establish the paths and locations these stars must take, given the truth of these observations.[28] And it proceeds a priori only in so far as in applying these theorems it derives these hypothetical necessities from mathematical essences as ontological grounds that determine their truth. Notice that sciences, like astronomy, that are in this way only impurely a priori presuppose for their very possibility purely a priori sciences, such as pure geometry, which deal solely with necessary truths.

Leibniz's account of real definition confirms that, in his view, the a priori knowledge of possibilities achieved in the demonstrative sciences is also knowledge from the ontological grounds of these possibilities, and thus consists in knowledge of what is essential, and so necessary, to these possibilities. A real, as against a nominal, definition specifies, not just any reciprocal property of what is defined, but a property that 'makes known the possibility' of what is defined.[29] Leibniz, moreover, contrasts a real definition that relies on experience for assurance that the nature it specifies is possible, with a real definition that itself reveals the possibility of what it defines, so as to prove this possibility a priori. The second sort of real definition he terms 'causal'. This terminology provides yet more evidence of Leibniz's adherence to the from-grounds notion of the a priori: the thought it reflects is that a real definition of something itself yields a priori knowledge of the possibility of what is defined in making clear how certain grounds determine its possibility, or nature. Indeed, he gives as an example of a causal definition one that 'contains the possible generation' of the thing defined.[30] More generally, a real definition makes the possibility of what it defines known a priori by resolving 'a notion into its requisites, that is into other notions known to be possible', where 'we know that there is nothing incompatible among them'.[31] It is in providing such definitions that, in Leibniz's view, reason establishes possibilities a priori, yielding certain knowledge that ideas are possible.[32]

[28] 'Finally, there are mixed propositions [i.e., propositions that mix propositions of fact and propositions of reason] which derive from premises some of which come from facts and observations while others are necessary propositions. These include a great many of the findings of geography and astronomy about the sphere of the earth and the paths of the stars, arrived at by combining the observations of travelers and astronomers with the theorems of geometry and arithmetic' (NE 446).

[29] DM § 24. [30] Ibid. [31] 'Meditations on Knowledge, Truth and Ideas'; AG 26.

[32] Cf. NE 446, cited above.

Leibniz terms a definition that is real and causal (and that thus proves the possibility of its object a priori) 'perfect' or 'essential' when it 'pushes the analysis back to the primitive notions without assuming anything requiring an a priori proof of its possibility'.[33] What Leibniz here calls 'the primitive notions' would seem, in turn, to be the 'primitive possibilities' or 'irresolvable notions' that he maintains are 'the absolute attributes of God'.[34] The fundamental instance of an essential definition, then, is the definition of God as the *ens perfectissimum*, 'the subject of all perfections',[35] where analysis of the notion of perfection reveals that perfections are simple, positive, and absolute qualities; since such qualities can be seen a priori to be compossible, this definition shows a priori, that is, from its ontological grounds, that an *ens perfectissimum* is really possible.[36] The essential definitions of particular contingent beings—which Leibniz held no finite mind could supply—would reveal their possibility from the ultimate ontological grounds, in the divine mind, of their possibility. For God, in Leibniz's view, is 'the source of all essence' and in such a way that the divine understanding contains ideas that express the essence of every possible being.[37]

In short, in Leibniz's view, if we restrict our attention to finite minds, a priori knowledge, or knowledge through a priori proofs, consists in rational perception of the necessity with which ontological grounds determine some truth, and thereby of why this truth obtains with universality and necessity. And what he terms 'a priori truths', or 'truths of reason', are necessary truths whose ontological grounds a finite mind can, in principle, appreciate as such through her reason. Moreover, in so far as a science is demonstrative, it proceeds by way of demonstrations that exhibit, with certainty, how the ontological grounds of a truth establish it as necessary (perhaps only *ex hypothesi*). This position, which takes the province of the demonstrative sciences to be providing explanations of necessary truths, by achieving insight into the grounds, or sources, of these truths, presupposes and develops the from-grounds notion of the a priori. The position persists, not only into the work of Leibniz's successors, such as Christian Wolff,[38] but also, as we will see, into that of Kant.

[33] DM § 24; AC 57. [34] AG 26. [35] L 167.

[36] For a detailed and highly illuminating discussion of this a priori proof, see Part II of Adams's *Leibniz*. [37] L 448.

[38] 'Most people know that the sun rises early in the morning from experience and cannot say why it happens. But the astronomer has insight into the grounds of the heavenly movements and the

In partial support of this last contention, I want to close the present section by pointing out how, in the *Nova Dilucidatio*, Kant operates with the from-grounds notion of the a priori, in the course of presenting a conception of a ground, or reason, common among his Leibnizian predecessors and contemporaries. In doing so, he does not, to be sure, use the term 'a priori'. But he none the less operates with what amounts to the traditional distinction between a priori and a posteriori grounds, so where 'a priori' and 'a posteriori' are understood in the from-grounds and from-consequence senses:

> To determine is to posit a predicate with the exclusion of its opposite. That which determines a subject in respect of a certain predicate is called the *reason*. Reason is differentiated into that which determines antecedently and that which determines consequently. That is antecedently determining the notion of which precedes that which is determined, that is to say, when it is not supplied, the determined thing is not intelligible. That is consequently determining which would not be posited unless a notion which is determined by itself had already been posited from some other source. You may call the former the reason why or the reason of being or becoming, the latter the reason that or the reason of cognition.[39]

Earlier, Kant has explained that, following Crucius, he uses 'determining ground' in a sense in which a determining ground is to be distinguished from a sufficient one; a sufficient ground may suffice to bring about a certain outcome without, as does a determining ground, excluding all alternatives to that outcome: 'The expression ''sufficient ground'' is ambiguous, as Crucius has adequately demonstrated, since it isn't at once clear how far it suffices; ''Determining'', however, means *a positing that excludes all alternatives*, and denotes that which definitely suffices for the thing to be grasped in one way only.'[40] Particularly telling is Kant's glossing, at 1: 396, 'antecedently determining reason' with 'the reason *why* [*rationem cur*]', and 'consequently determining reason' with 'the reason *that* [*rationem*

connection of the earth with the heavens, and knows the same thing through reason, he can demonstrate that, why, and at what time it *must* happen' (*Deutschen Metaphysik*, § 372; cf. § 77). Des Hogan cites these passages in the course of a valuable discussion of the rationalisms of Kant's predecessors in 'Three Kinds of Rationalism and the Non-Spatiality of Things in Themselves' forthcoming in the *Journal of the History of Philosophy*.

[39] 1: 396.

[40] 1: 393. Here I am indebted to Des Hogan's helpful discussion of this distinction between sufficient and determining grounds in his 'Three Kinds of Rationalism'; Hogan traces the origin of Kant's distinction in Crucius, focusing on its role in their accounts of the freedom of the will.

quod]'. Recall that the essential difference between a priori and a posteriori grounds, for Leibniz and his successors, is that the former do, and the latter do not, explain what they ground. Furthermore, the explanation in question is grasped in appreciating a determining ground as such, and so how that ground excludes every other alternative, rendering this consequence necessary. This shows that Kant's distinction between antecedently and consequently determining grounds is tantamount to Leibniz's distinction between a priori and a posteriori grounds.

Kant puts as appositive to 'the ground why', 'the ground of being [*rationem essendi*] or becoming [*fiendi*]', and to 'the ground that', 'the ground of cognition [*rationem cognoscendi*]'.[41] But, in doing so, he clearly does not mean to suggest that antecedently determining grounds cannot also serve as grounds of cognition. Consider how he goes on to illustrate the distinction using the example of the fact that successive propagation of light takes place at an assignable velocity. The eclipses of the satellites of Jupiter are consequently determining grounds of this fact: in the order of being, that these eclipses occur is a consequence, and not a ground, of this successive propagation. On the assumption that Descartes was correct, he tells us, the antecedently determining ground of this fact is the elasticity of the elastic globules of ether: in the order of being, this elasticity is the ontological ground of the fact that light is propagated successively at an assignable velocity. This example illustrates how, in Kant's view, an antecedently determining ground can serve, not only as the ground of a fact, but also as the ground for our knowledge of that fact. And it shows that the point of the appositives is to indicate that a consequently determining ground, unlike an antecedently determining ground, is as such *merely* a ground of cognition, and *not* a ground of being. Moreover, the example confirms that what distinguishes antecedently and consequently determining grounds is being prior or posterior in the order of being: in particular, what makes something a consequently determining ground is its being a consequence, in the order of being, of what is, in the order of our cognition, the ground. It thus confirms that this distinction is one between a priori and a posteriori grounds, in the traditional senses of 'a priori' and 'a posteriori'.

The distinctions among different kinds of ground that Kant draws in the *Nova Dilucidatio* are common currency within the Leibnizian tradition.

[41] 1: 396.

Wolff, Baumgarten, and Crucius—all distinguish grounds of cognition from two species of antecedently determining grounds, grounds of becoming (*ratio fiendi*), and so of the actual existence of some contingent being, and grounds of being (*ratio essendi*), that is, grounds of something's possibility.[42] Since only contingent beings can come to be, only these beings, and truths of fact (which concern these beings), have *rationes fiendi*. The actual existence of the necessary and eternal being, God, does have a ground, but that ground is God's *ratio essendi*—God's essence so that God's very possibility entails God's actual existence—and for this very reason God is a necessary being and cannot have a *ratio fiendi*.[43] Particularly noteworthy is the fact that Baumgarten identifies essence both as the principle of being and as the principle of the cognition of modality.[44] This indicates that he, with Leibniz, regards essence, the subject-matter of an essential definition, as the fundamental *ratio essendi* and so the proper ground for the cognition of modality.

This distinction between *rationes essendi* and *rationes fiendi*, and the conception of essence as the *principium essendi*, will prove important to understanding Kant's account of the a priori cognition of a finite mind. As we will see, in this account a finite mind cognizes purely a priori only in achieving a rational perception of *rationes essendi* as such. Such a perception, moreover, relates *rationes essendi* to an essence: an essence specifies the inner possibility of some thing as that in which all its essential properties are united, and it is as a constituent contained in something's essence that a property constitutes an essential property of a thing, and thus a *ratio essendi* of that thing. Moreover, it would take an infinite mind to cognize

[42] Wolff, *Prima Philosophiae sive Ontologia*, § 874; Baumgarten, *Metaphysica*, § 311; Crucius, *Entwurf der nothwendigen Vernunft-Wahrheiten*, § 34. For an excellent discussion of these figures' conceptions of ground, and their relation to the pre-critical Kant's conception of ground, see the second chapter of Eric Watkin's *Kant and the Metaphysics of Causality* (New York: Cambridge University Press, 2005).

[43] The *rationes essendi* of contingent beings, traditionally held to be cognitions contained in God's understanding, explain only the possibility of these beings; their actual existence is to be explained by their *rationes fiendi*, contained in the will of God. For an illumining discussion of these points, see John Carriero's 'Leibniz on Infinite Resolution and Intra-Mundane Contingency', Parts I and II. I believe these points have important implications for Kant's distinction between theoretical and practical reason/philosophy.

[44] '*Essentia est principium essendi et cognoscendi modorum*' [sive 65, 50] (*Metaphysica*, § 311). Baumgarten's singling out the *ratio essendi* as what provides grounds for cognition of modality implies that the *ratio fiendi*, the ontological ground of some thing's coming to be, and so of the actual existence of a contingent being, does not yield cognition of necessity or possibility. This remark of Baumgarten's also indicates that he, too, holds that a *ratio essendi*, and so a real ground, can also serve as a *ratio cognoscendi*.

contingent beings purely a priori through their *rationes fiendi*: a finite mind can register *rationes fiendi*, at best, only in mixed propositions. In Kant's terminology—which reserves the term 'cause' (*causa, Ursache*) for a *ratio fiendi*[45]—reason (which, being discursive, is a capacity of cognition had by a finite mind) cannot discern the causality of causes. In these respects, Kant's account of the a priori retains positions carved out by Leibniz, and retained by his followers. Indeed, what motivates his critical turn are misgivings about the extent to which we can cognize things purely a priori. These misgivings lead him finally to the hypothesis that the only essences, and *rationes essendi*, that we can achieve rational perception of are, in the first instance, grounds of the possibility, not of things, but rather of our cognition of things.

2. Kant's Explicit Use of the From-grounds Notion of the A Priori

I turn now to providing an initial case for reading Kant as operating in the first critique with the from-grounds notion of the a priori. In the present section, I will focus on the opening paragraphs of the *Metaphysik Mongrovius*, a student's notes on the lectures on metaphysics Kant held in 1782–3. This passage is of interest not only because it contains explicit expressions of the from-grounds notion of the a priori but because it puts this notion to work in providing a general overview of the new, transcendental approach to metaphysics that Kant pursues in the *Critique of Pure Reason*. In this approach, the metaphysician aims to exhibit in purely a priori cognition what are, in the first instance, the antecedently determining grounds not of things themselves but rather of our cognition of things. Now this cognition is cognition, in respect of its form, of our experience in general and so of the *ratio essendi* of experience in general. Moreover, there is nothing more to the being of the objects of our possible experience than their being objects of our experience in general. So, in exhibiting, in a priori cognition, the real formal ground of our cognition of objects, the critical

[45] 'Cause and ground are to be distinguished. What contains the ground of possibility is ground [*ratio*], the principle of being [*principium essendi*]. The ground of actuality is the principle of becoming [*principium fiendi*], cause [*causa*]. What contains the ground of something is called in general principle [*principium*]' (28: 571; cf. 572).

metaphysician, at one stroke, exhibits as such the *ratio essendi* both of these objects themselves and of all our a priori cognition of these objects (be it the a priori cognition of mathematics, physics, or transcendental philosophy).

Consider first that, according to Mongrovius, Kant explicitly characterizes the a priori and the a posteriori in their original senses: 'If I begin from the consequences, then I cognize something a posteriori; if I begin from the grounds, then I cognize a priori'.[46] Kant then infers from this characterization of the a posteriori the now-familiar characterization in terms of dependence on experience: 'Cognition taken from experience is eminently [*per kat' exochen*] a posteriori, and from now on when we call cognitions a posteriori, then we are always understanding these to be from experience, because experience contains the last consequence of our cognition, for which we seek grounds by means of reason'.[47] We will return to these texts, and to the larger passage in which they are to be found, shortly. But, for the present, note simply that Kant does not define the a posteriori in terms of experience, but rather infers from the from-consequences characterization of the a posteriori, together with an account of experience, that cognition 'taken from experience' is a posteriori, and eminently so.

Despite the fact that these texts are drawn from a student's lecture notes, they provide important evidence in support of ascribing the from-grounds notion of the a priori to the Critical Kant. To begin with, the context makes it abundantly clear that Kant is not merely explicating notions employed by Baumgarten in his *Metaphysics* (the class text) notions to which he does not, himself, subscribe. Kant has not even turned to discussing the *Metaphysics*. He is rather clearly—and at points entirely explicitly (cf., e.g., 29: 752–3 where he refers to the approach he has been presenting as transcendental philosophy)—providing a detailed account of his own, distinctively critical, approach to metaphysics, and, in this context, is putting these notions of the a priori and the a posteriori to work in articulating his own views about human cognition and its transcendental conditions. Moreover, the *Metaphysik Mongrovius* is an especially detailed and significant set of student notes whose general reliability can be confirmed by comparing its contents to other good notes from the 80s.[48] In any case, the present and following

[46] 29: 748. [47] Ibid.

[48] See Karl Ameriks and Steve Naragon's Introduction to *Lectures on Metaphysics* (New York: Cambridge University Press, 1997).

sections will provide textual evidence from central published critical texts to corroborate what is drawn from these lecture notes.

Now, as mentioned above, Kant does occasionally speak of knowing (*Wissen*) something a priori. But, more commonly what Kant talks about in the first critique (and more generally, throughout his critical philosophy) is cognition (*Erkenntnis*). Most commentators have, in effect, treated *wissen* and *erkennen* as equivalent. However, as I have argued at length elsewhere, distinguishing them is crucial to understanding the sense and motivation of Kant's critical philosophy, and in a myriad of different ways.[49] In the present context, this distinction is important to note because most of the textual basis for ascribing the from-grounds notion of the a priori to Kant—including the passage just cited from the *Metaphysik Mongrovius*—concerns a priori cognition, and can be understood properly only if one attends to his notion of cognition. My reading of this notion will, itself, be controversial, and I will not be able to provide a full development, let alone an adequate defense, of it here. However, having some sense of what Kant means by 'cognition' is essential, not only to recognizing that he retains the from-grounds notion of the a priori, but also to seeing how he puts this notion to work in developing and executing the project of the first critique.

What, then, does Kant mean by cognition, or cognizing? We read in the Jäsche *Logic* that, in cognizing a thing, we represent it with consciousness, in respect of its identity, as well as its diversity, in comparison to other things.[50] The content of an act of cognizing—whatever one consciously represents in cognizing, considered in so far as one so represents it—is a cognition. Moreover, if we are to cognize a thing, as against merely think it, we must be able to prove the real possibility of that thing (Bxxvi*n*). So, for

[49] 'What Can We Know about Things in Themselves?' (unpublished MS). For Kant, *erkennen* and *wissen*, though distinct, are closely related: *wissen* of something is not just any assent with the requisite degree of certainty, but rather such assent that is based on cognition of the same; indeed, *Wissen*, in Kant's sense, consists in *Erkenntnisse* in so far as we cognize the systematic unity they have under principles. In the above-mentioned piece I develop in detail the reading of Kant's notion of cognition that I will go on to sketch below, in the service of clarifying the grounds on which Kant denies that we can have cognition of things in themselves and showing that both the sense, and the grounds, of this denial are consistent with Kant's laying claim to the knowledge, in our familiar sense, regarding things in themselves advanced in his transcendental account of our theoretical cognition of things.

[50] More precisely, Kant characterizes *erkennen* as '*kennen* with consciousness', having characterized *kennen* as 'to represent something in comparison to other things in respect of identity as well as diversity' (9: 64–5). These characterizations concern specifically the objective content of our cognition; they are not intended to hold for the self-cognition articulated in logic or transcendental philosophy, and so for all our cognition.

example, Kant counts our concept of gold as a cognition. For, in grasping this concept, we consciously represent certain predicates (malleability, density, etc.) as grounds that determine certain objects presented to us in our experience in respect of their specific identity. And, according to Kant, we can prove the possibility of gold by appeal to our experience of gold, since these experiences show that gold actually exists, and whatever is actual is possible.[51]

We cognize something in respect of its specific identity, then, when in the requisite fashion we consciously represent a general predicate as one that is essential to it. But, according to Kant, we can also cognize something in consciously representing it in respect of its numerical identity, as well as its diversity, in relation to other things. We do so when we cognize a thing in our sensible intuition. What this cognizing amounts to is, admittedly, a difficult exegetical matter. I am inclined to think that, in Kant's view, we cognize a thing when we perceive it directly in space and time as a persisting individual throughout a perceptually given spatio-temporal manifold of things, relative to the other things given in that manifold. For we thereby represent a phenomenon, with consciousness, in respect of what individuates it as a phenomenon—namely, the determinate spatio-temporal position it has, as a persisting object, within the entire phenomenal world. Moreover, our cognition in intuition, like our cognition through concepts, consists of predicates, or properties, in so far as we are conscious of them in such a way that they provide us grounds for positive determinate judgments of a thing in respect of its identity.[52]

If we restrict our attention to the first sort of cognition—cognition in respect of specific identity—we can characterize cognition as conscious content through which the subject of that consciousness can understand something. But, for Kant, this is not a correct characterization of the second sort of cognition, and so for cognition in general: in representing a thing, with consciousness, in one's sensible intuition, in respect of its numerical identity, one does not thereby *understand* that thing in respect of its *distinctive*

[51] Note how Kant's requiring that cognition have an object whose real possibility the subject of that cognition can prove echoes Leibniz's discussion of real definition (cf. section 1 above).

[52] Kant terms identifying properties, as they provide grounds of cognition to a conceptual understanding, a mark (*Merkmal*). Marks, as they constitute grounds for cognizing in respect of numerical identity, he terms singular, or intuitive, marks, and as they constitute grounds for specific identification, general marks. I develop this reading of marks, intuitive and discursive, in 'Kant on Marks and the Immediacy of Intuition', *Philosophical Review* (2000).

numerical identity. Indeed, Kant denies that there are singular concepts: he specifies generality as a defining property of concepts,[53] and, invoking the law of continuity, he asserts that there cannot be a lowest species.[54] Since our understandings are conceptual, or discursive, we cannot understand a singular object, as such (although we can understand, in general, what it is for an object of our cognition to be a singular object).

Here, then, is a first sketch of Kant's contrast between cognizing something from its grounds, or a priori, and cognizing it from its consequences, or a posteriori. Since to cognize something is to be conscious of it in respect of its identity, to cognize it from its grounds is to be conscious of its identity from a rational perception, as such, of the ontological grounds that determine this identity, as they are sufficient to determine this identity. So, for example, we cognize a triangle a priori in cognizing how being three sided, enclosing a space, and being single-planed suffice collectively as ontological grounds to determine the specific identity of a figure as a triangle. In Kant's account, we cognize a triangle a priori in constructing the purely sensible concept *triangle*, that is, in exhibiting in pure intuition a figure corresponding to this concept.[55] Moreover, this construction constitutes a real definition, one in which reason establishes the real possibility of its object a priori, because it generates the schema of this pure sensible concept, a rule of imagination that expresses the possible generation of the figure.[56] To cognize something a posteriori, in contrast, is to be conscious of its identity, not from such a rational appreciation of the ontological grounds that determine its identity, as they suffice to determine this identity, but rather merely from the perception of what one takes to be its effects. This, according to Kant, is how we cognize gold when we form the concept of gold from our experiences of gold.

The present reading of Kant's conception of a priori cognition, and more generally my ascribing the from-grounds notion of the a priori to him, finds confirmation in a gloss on cognizing something a priori that Kant offers in the Preface to the *Metaphysical Foundations of Natural Science*. In particular, Kant glosses our cognizing something a priori with our cognizing it 'from its mere possibility'.[57] Now what determine the mere possibility, as against the actuality, of something are its *rationes essendi*, as

[53] A377/B320; cf. 9: 91. [54] 9: 97–8. [55] A713/B741–2.
[56] A729–32/B757–60; A140/B179–80. [57] 4: 470.

they collectively constitute its essence or nature. For example, according to Kant, we cognize a triangle a priori in doing Euclidian geometry through our act of constructing the concept of that figure, an act in which we consciously follow certain principles to exhibit, in a priori intuition, an object corresponding to that concept.[58] Our cognizing a thing (something that is, as the subject of power, real) a priori would, in an analogous vein, consist in our cognizing it from its nature, in the formal sense—that is, 'the first inner principle of everything that belongs to the being [*Dasein*] of a thing'[59]—in determining the dynamical constitution that constitutes its being as a thing. For example, to cognize gold a priori would be to cognize the necessity with which its malleability, density, etc. follow from its nature. We cognize something a priori, then, when we cognize it in and through a rational perception of the way in which its nature or essence determines, so as to necessitate, its essential properties. In Kant's terminology, when we understand something from its *ratio essendi* in this way, we have insight into it.[60]

Note that cognizing gold a priori is a higher cognitive achievement than merely understanding gold. For, as we saw earlier, Kant holds that we can understand gold only by framing the empirical concept of a solid, dense, yellow, and malleable body. Indeed, Kant agrees with Hume that we cannot cognize a priori any object of our experience in respect of its distinctive empirically given character. So, for example, Kant holds that we cannot achieve insight into the nature of gold. That is to say, we cannot achieve rational perception of the way in which the nature of gold, as the first inner principle of all that belongs to its being, necessitates gold's being a solid, dense, yellow, malleable body. More generally, Kant holds that, *pace* Leibniz, no conceptual understanding can achieve insight into specifically distinct natures of the objects of its cognition.[61]

[58] A713/B741–2. [59] IV, 467.

[60] See the Jäsche *Logic*, 9: 64–5. Kant also treats cognizing a priori as equivalent to having insight at A737/B765, cf. A760/B788. And, in section V of the published Introduction to the *Critique of Judgment*, Kant glosses what it is for something to be 'contingent for our insight' with its being such that we cannot cognize it a priori (5: 183).

[61] 5: 183. Kant does maintain, in the *Metaphysical Foundations of Natural Science*, that we can have insight into how the transcendental principles of our cognition, together with mathematics, determines certain fundamental laws of physics—such as Newton's law of universal gravitation. In his account, such universal laws of nature yield a 'thoroughgoing interconnection of empirical cognitions into a whole of experience'; but this is an interconnection 'among things with respect to their genera, as things of nature in general, but not specifically, as such and such particular beings in nature' (5: 183).

It will be important to see that any conscious content through which one can identify something counts as a cognition, whether or not in that consciousness itself one represents that content as a ground that determines this identity. In particular, in having an a priori cognition one need not be conscious as such, of an ontological ground of something's identity, so as actually to cognize something a priori through that ground. It is enough that one can, perhaps only together with other a priori cognitions, achieve insight through that consciousness. For example, we can actually possess the categories, or pure concepts of the understanding, without thereby having any insight into anything. None the less, these concepts constitute a priori cognition in so far as we can, by grasping how these concepts are functions essential to realizing the original synthetic unity of apperception, come through these concepts to have insight into the nature of our capacity of understanding and its possible objects.

With this sketch of what Kant means by cognizing a priori and a posteriori in hand, let's return to the *Metaphysik Mongrovius*. The lecture begins with the claim that 'our cognitions are in composition [*im Zusammenhang*] in a two-fold way'.[62] The first Kant describes 'as an aggregate, when one is added to another to constitute a whole, e.g., a sand hill is not in itself a connection of things, but rather they are arbitrarily put together'.[63] The second he describes as 'as a series of grounds and consequences, the parts of the series called members because we can cognize one part only through the others'.[64] Moreover, he remarks that these grounds and consequences make a connection (*Verknüpfung*) 'according to a rule', and implies that the second composition thereby provides determinate concepts of the whole, in a way the first manifestly cannot.[65]

Later in the *Metaphysik Mongrovius* Kant elaborates on what he means by 'connection'. He begins his discussion of ground and consequence by remarking, 'Ground (*ratio*), relation (*respectus*), is a manifold [of elements] insofar as one is posited or canceled by another. Thus all relation is relation either of connection or of opposition (*relatio vel nexus, vel oppositionis*)'.[66] He then specifies that 'The relation of ground and consequence and vice

So, for example, Newton's insight into the law of universal gravitation does not itself constitute, or in any way require, insight even into the nature of the object of our empirical concept of matter in general. For present purposes, I set to one side Kant's discussion of chemistry in the *Opus Postumum*.

[62] 29: 747. [63] Ibid. [64] Ibid. [65] Ibid.
[66] 29: 206.

versa is connection (*nexus*)'.[67] This specification tells us two things. First, a manifold of elements makes up a nexus in so far as positing one element requires positing another. Second, as the term 'vice versa' indicates, a connection consists not only of the relation of ground to consequence, but also of consequence to ground. After all, positing a consequence requires positing a ground—since every consequence, as such, must have a ground. But what, then, is the difference between a ground and a consequence, given that positing a consequence also requires positing its ground? Kant's answer is implicit in the following definition: 'The ground is that which, having been posited, another thing is posited *determinately*'.[68] He specifies, moreover, that positing determinately amounts to positing 'according to a general rule', and he infers from the ground's giving this rule that 'the connection of the ground and the consequence is necessary'.[69] In characterizing relation (*respectus*), Kant singles out ground, as against consequence, because the ground is what, fundamentally, constitutes relation, be it a relation of connection or opposition.

The distinction between two types of composition with which Kant opens the *Metaphysik Mongrovius* relates interestingly to a distinction he draws in the B-Edition of the first critique between two types of synthesis, or combination (*Verbindung*): composition (*Zusammensetzung*), which is a relation of homogenous elements that do not, themselves, belong together necessarily, and connection (*Verknüpfung*), a relation of heterogenous elements that do.[70] The former type of synthesis is governed by principles of pure understanding (i.e., those of the axioms of intuition and the anticipations of perception) that Kant terms 'mathematical' to indicate that they make mathematical cognition possible.[71] It is only in so far as we direct this synthesis—in acts of mathematical construction, to generate determinate quantities according to mathematical concepts and axioms—that we generate a connection, a *nexus*, among the elements of the homogenous manifold contained in the forms of our sensible intuition. The second sort

[67] Ibid. [68] 29: 208; italics mine.

[69] Ibid. In the course of providing this elaboration of what he means by 'connection', Kant once again articulates the from-grounds notion of the a priori. He writes, 'The thing as ground relates a posteriori to its consequence, i.e., connection (*nexus*), and the reverse is a priori connection (*nexus*)' (29: 807). The point of the first clause is that a thing, in so far as it determines a consequence, is cognized a posteriori from that consequence, a relation that is a connection in the sense he specifies. The point of the second, is that a consequence is cognized a priori from the thing that is its ground, and that this relation is an a priori connection. [70] B201 n.

[71] Ibid.

of synthesis, in contrast, is governed by the dynamical principles (those of the analogies of experience and the postulates of empirical thought).[72] Moreover, the opening of the *Metaphysik Mongrovius* echoes Kant's claim, in the A-Edition Transcendental Deduction of the Categories, that unless the appearances apprehended by the imagination conformed to rules of association and the principle of the synthetic unity of apperception, they would make up a mere heap and lack the objectively necessary combination they require to constitute cognition.[73]

These points suggest that the series of cognitions in question in the *Metaphysik Mongrovius* consists, most fundamentally, of relations of ontological ground and consequence that our cognitions stand in in so far as they do, or can, constitute a single universal experience. And of particular importance are the relations of ontological ground and consequence that particular given appearances, the material of our experience, stand in in so far as they, under the principle of the synthetic unity of apperception, realize the form of our experience to constitute experience: these relations of ontological ground and consequence are the determinate causal relations that particular objects of experience must, as such, stand in according to the principles of the Analogies of Experience. The principle of the Second Analogy, for example, specifies that appearances constitute cognition only in so far as they designate substances that stand, under particular empirical laws, in relations of cause and effect: our experience of the sun and our experience of wax, as members of this series, stand in the relation of cause and effect, in so far as the sun is, under a particular empirical law of nature, the cause of the wax's melting.

In the opening passage of the *Metaphysik Mongrovius*, Kant turns next to expanding on the connection of our cognitions in a series of grounds and consequences. He compares this connection, in virtue of which our cognitions constitute cognitions, to the connection things that are members of a series of grounds and consequences have as members of this series:

> With grounds and consequences we must think of a priori boundaries, i.e., a ground that is not also a consequence, and a posteriori boundaries, i.e., a consequence that is not a ground, e.g., with human generations: human beings are members in a series, yet here we must think of a human being who does conceive, but is not born, thus an a priori limit (*terminus*), and of one who is born but conceives no one,

[72] B201 n. [73] A121–2; cf. A104–5, A108–9, and A112–13.

thus an a posteriori limit (*terminus*). We consider here (in metaphysics) not things as they are connected as grounds and consequences, but rather cognitions, which also have a descent like human beings or other things. I can imagine a cognition that is not a consequence, thus the highest ground, and one that is not a ground, thus the last consequence. {The last consequence is an immediate experience, e.g., something-body-stone-limestone-marble-marble column.} We thus have an idea of a connection of cognitions as grounds and consequences.

In contrasting a priori and a posteriori limits of a series of things, Kant clearly uses 'a priori' and 'a posteriori' in the from-grounds and from-consequences senses: what makes some thing an a priori limit of a series, for example, is its being a member of that series that is an ontological ground of other members, and an ontological consequence of none. Moreover, when he turns to the series of grounds and consequences formed by cognitions, he specifies that such a series, as against a series of things, is the subject-matter of metaphysics. This specification indicates that, as I claimed earlier, Kant is here sketching his own, distinctively transcendental, approach to metaphysics: for what distinguishes this approach is that it seeks, in the first instance, insight not into the ontological grounds of the possibility of things themselves, as traditional rationalist metaphysics did, but rather into the ontological grounds of the possibility of our cognition of things. As we will see, metaphysics, in this approach, takes special notice of the cognitions that are the a priori and a posteriori limits of the series that our cognitions form.

Kant continues by introducing the idea of an a posteriori principle, which he characterizes as a principle of cognition, and not a principle of being. This passage deserves close scrutiny, not only because the notions of the a priori and the a posteriori it invokes are clearly the from-grounds and from-consequences ones, but because Kant will shortly characterize metaphysics as 'the science of the a priori principles of human cognition':[74]

Cognitions which are the grounds of grounds that follow a certain [*gewissen*] rule are called principles (*principia*). Thus insofar as cognitions are in a series, there must also be principles (*principia*). *It is remarkable that I can make a consequence into a ground, but one out of which* the other does not [follow], but rather through which I always arrive at the cognition of the other. Thus they are not principles of being (*principia essendi*) but of cognition (*cognoscendi*), e.g., I cognize God's being

[74] 29: 749.

[*Dasein*] from the world. The world is nevertheless not the ground of God, but rather the reverse, but through the world I am able to arrive at the concept of God, and to this extent I can move from the principled (*principiis*) to the principles (*principia*)—the consequences that are used as grounds, for going back in reverse to their own grounds, are called a posteriori principles (*principia*). If I begin from the consequences, then I cognize something a posteriori; if I begin from the grounds, then I cognize a priori. If something is given to me then I can test whether I could indeed have cognized it a priori from grounds, e.g., experience teaches that sunlight melts ice, but we would have hardly cognized this a priori.[75]

A ground that follows a certain rule is a ground from which a determinate consequence follows according to a general rule. Any cognition consists of such a ground, as it is registered as such in a subject's consciousness, through which that subject may determine something in respect of its identity. A cognition that is the ground of such grounds—what Kant says is called 'a principle'—is evidently the content of an act of cognizing that, as such, is the ground of 'grounds that follow a certain rule' constituting such grounds. Kant's position seems to be that any ground requires a principle, a ground that makes it a ground: as we will see, if this thought is not to lead to a vicious regress there must be first principles, grounds that ground themselves as grounds. Since cognitions constitute grounds in so far as they comprise a series, there must be principles that ground cognitions' comprising a series.

This reading, at any rate, accounts for Kant's inferring the second sentence of this passage from the first. Moreover, this reading fits nicely with what he goes on to say in the next two sentences. In the italicized sentence, he claims that, remarkably, we can make an ontological consequence into a ground of the cognition of its ontological ground. He then infers from this claim that the principles we thereby employ are principles of cognition, not principles of being.

Kant turns next to providing an illustration, drawn from Leibnizian rational cosmology, of the distinction between a principle of being and a principle of cognition, an illustration that also allows him to introduce the idea of an a posteriori principle without appealing to the notion of independence from experience. According to this cosmology, God is the principle of being of the intelligible world: God, through his self-cognition, constitutes the ground of the possibility of a substantial

[75] 29: 748.

whole of things standing in a nexus of real connection.[76] But the world, so understood, is the principle of my cognition of God as the ground of the possibility of the world: in particular, it is our cognition of the world as, in respect of its possibility, an ontological consequence of God that makes the world a ground of our cognition of God as the being that is the ultimate ground of all possibility. And since what makes the world a ground of cognition of God is its being, in respect of its possibility, an ontological consequence of God, the world is an a posteriori principle in the sense Kant specifies: the world is an ontological consequence that is used as a ground of cognition for going back to its ontological ground. It is now clear that, in the ensuing characterizations of cognizing a priori and a posteriori, cited at the outset of the present section, Kant refers to ontological grounds and consequences, and thus that he is offering the traditional, from-grounds and from-consequences characterizations of the a posteriori and the a priori.

There are two points about Kant's use of this illustration that are important to see. The first is that he does not mean to be endorsing either the rational cosmologist's idea of the intelligible world or her use of this idea in establishing God as the ground of all possibility. In Kant's own view, we cannot, on theoretical grounds, prove the real possibility of the intelligible world (things in themselves as we conceive of them through the ideas of pure reason). We can prove, on theoretical grounds, the real possibility only of the sensible world, which is the series of cognitions we form out of sensible impressions in interpreting appearances as constituting experience. For this reason, experience is the only genuine a posteriori principle of our cognition, and the a posteriori is properly lined up with the empirical. Second, the cognition (or, better, the problematic concept) of God that the rational cosmologist generates in taking the world as an a posteriori principle includes the idea of God as an *ens per se*. For God can serve as the principle of possibility that is the a priori limit of the series that constitutes the intelligible world only as a being that does not, in any way, depend for its being on any other being. The idea of God as an *ens per se*, however, is misunderstood, if it is taken as a positive characterization of God's being (*Dasein*). That would be to conceive of God as *causing* God's own existence, and this, Kant stresses, is incoherent. The idea of God as an *ens per se* is, rather, a merely relative characterization of God: it says only

[76] Cf. 29: 849–50.

that God exists independently of any other being, and does not characterize positively how God has God's being.

Having illustrated his notion of an a posteriori principle with the case of the rational cosmologist's idea of the intelligible world, Kant turns to what, in his critical view, he holds are the only a posteriori principles for our genuine cognition (as against mere thought): experience and the universal a posteriori cognitions that can be drawn, through induction, from experience. Having remarked that the universal cognitions experience teaches us (such as that sunlight melts ice) are ones that we 'would hardly have cognized a priori', he continues:

> Cognition that is taken from experience [is] eminently [*per kat exochen*] a posteriori, and from now on when we call cognitions a posteriori, then we are always understanding these to be from experience, because experience contains the last consequence of our cognition, for which we seek grounds by means of reason. If we take experience as the principle [*principium*] then the principle is empirical: e.g., experience teaches all bodies are heavy (insofar as we are acquainted with them), we can accept it as a principle and say: since all bodies are heavy, it follows that . . . A priori principles are those which are not borrowed from any experience. Whether there are such must be investigated shortly.

For our purposes, the crucial thing to see is how experience is a principle, and an a posteriori one at that, in the sense of 'principle' Kant has specified. Experience is cognition of a thing in empirical intuition, and as such consists of a manifold of appearances that provides a subject grounds for cognizing something in respect of its numerical identity. But a composition of appearances provides such grounds, and so constitutes experience, only if the subject can cognize this composition as a connection: that is, as an instance of a regularity in kind among appearances that manifests a power, one that constitutes the specific identity of the individual thing that is, in virtue of affecting our sensibility to produce these appearances, the object of experience. For it is only in exercising token instances of such powers in mutual interaction with one another that things occupy the determinate spatio-temporal locations, as persisting objects, that individuate them as phenomena. Now, in Kant's account, experience itself consists only in the mere composition of appearances, and contains no consciousness of necessary relations among appearances.[77] What experience teaches us,

[77] A176/B219.

strictly, is thus only what we have observed, e.g., that sunlight has melted ice on all the occasions we have witnessed. When we take such a teaching of experience as a principle, we are not merely taking these regularities among appearances we have observed holding universally—at all times and places—as an ontological consequence of the grounds these appearances have in the things that affect our sensibility to produce these appearances in us. We are thereby taking these regularities to be, as such, an ontological consequence that is a ground of a posteriori cognition of those things. And we are thereby taking experience as a principle, because we are taking experience, in so far as it contains appearances (the last consequence of our cognition), to be the ground that makes these appearances 'grounds that follow a certain rule': experience is the principle that makes appearances of the sun and ice a posteriori cognitions of the sun and ice.

We are now in a position to see what an a priori principle is, in Kant's fundamental sense. An a posteriori principle is a cognition that is the ground of determining grounds as an ontological consequence. An a priori principle, then, is a cognition that is the ground of determining grounds as an ontological ground. Moreover, Kant implies at 29:748 that an a priori principle of our cognition is both a principle of cognition and a principle of being (*principia essendi*). This suggests that our a priori principles are cognitions that are grounds of determining grounds as principles of the possibility of those grounds. Now, in the last cited passage, Kant characterizes a priori principles as ones that are not borrowed from any experience. This reflects his having just specified that in calling cognitions 'a posteriori cognition' he will always understand them as cognitions that are borrowed from experience. What motivates this specification, moreover, is Kant's claim that the last consequence of our cognition consists of appearances insofar as they are contained in experience: the characterization of the a posteriori in terms of experience has the virtue of focusing our attention on the a posteriori principles of our genuine cognition. The corresponding characterization of a priori principles shares the same virtue. None the less, the characterizations of the a posteriori and a priori in terms of experience clearly derive, in Kant's view, from the more fundamental, from-consequences and from-grounds ones.

Recognizing that Kant is working with the original, from-grounds and from-consequences notions of the a priori and the a posteriori sheds light on Kant's claim, a little later in the *Metaphysik Mongrovius*, that in determining

the boundaries of metaphysics, we need to recognize that a priori principles do, and a posteriori principles do not, 'carry necessity with them'.[78] All Kant says by way of elaborating these claims is to illustrate how a posteriori principles do not carry necessity with them: he tells us that experience can at best teach us only that every event we have observed has had a cause; 'it cannot deny that something might sometime occur without a ground'.[79] He seems to take it as evident from what he has said so far how a priori principles carry necessity with them. We can now see why he does. An a priori principle of our cognition is such a principle as an ontological ground. What is more, it is such a principle not just as any ontological ground, but specifically as a *ratio essendi* which is registered as such in a rational perception. It thus carries necessity with it: for this rational perception is one of the necessity with which such grounds rule out, as impossible, any exceptions to this principle.[80]

Indeed, in light of what we have seen Kant means by '*principium*', we can now see how, in important respects, his account of the a priori principles of human cognition is continuous with an account common within the Leibnizian tradition. In particular, we can see that, as I claimed at the end of Section I, Kant follows Baumgarten in holding that essence is the fundamental ground of *rationes essendi*. When Baumgarten claims that '*Essentia est principium essendi et cognoscendi modorum*',[81] he is claiming both that thing's essence is the *ratio essendi* of all its *rationes essendi* and that, as such, essence is the *ratio essendi* of all cognition of modality, cognition that is a priori in the from-grounds sense of 'a priori'. Kant endorses the position that essence is the principle of being when he remarks that the principle of being (*principium essendi*) 'concerns the essence [*Wesen*] of things'.[82] And he also endorses the view that *rationes essendi* are the grounds of cognition of modality when he characterizes a *ratio essendi* as a ground of that which belongs to a thing considered according to possibility.[83] For example, he tells us, when we consider a triangle according to its possibility, the three

[78] 29: 749. [79] Ibid.

[80] In this connection recall that, on the Leibnizian tradition, not all a priori cognition—cognition from ontological grounds, grasped as such—must be cognition of what is necessary. On Leibniz's own view, for instance, the divine understanding can cognize contingent truths a priori from their *rationes fiendi*. So there is reason to think that, in characterizing a priori principles in terms of necessity, Kant means only to be characterizing the a priori principles of our cognition, or perhaps more generally of the cognition had by a conceptual understanding. [81] *Metaphysica*, section 311.

[82] 29: 844. [83] 29: 809.

sides of a triangle are a *ratio essendi* of the three angles; if we consider a triangle in actuality, in contrast, we have to do with a *ratio fiendi*, such as ink and qull, and so with a cause.[84]

That an a priori principle of our cognition is an ontological ground of determinate grounds that is registered as such in a rational perception invites the question Kant poses at the end of the last cited passage: are there any a priori principles? Indeed, in the remainder of the introduction to metaphysics he provides in the *Metaphysik Mongrovius*, he argues that this question is made pressing by the abysmal failure of metaphysics. The failure of ontology, as well as of rational psychology, rational theology, or rational cosmology, to provide any genuine insight into the nature, either of being in general, or of the soul, God, or the world calls into question whether there are any a priori principles at all, including the principles of mathematics. Solving this problem, he claims, will require a metaphysics that employs an entirely new method, one that begins by shifting our attention from the objects of our reason, to reason itself, and focuses on cognizing the possibility of pure reason purely a priori. And he names this initial, purely a priori, part of the new metaphysics—one that corresponds to what has wrongly been termed 'ontology' or 'general metaphysics'—'transcendental philosophy'.[85] In all this, Kant is clearly reviewing, in terminology that enjoyed currency within the Leibnizian tradition, the account of the problem of pure reason, and the proposal to solve this problem by executing a tribunal of pure reason that he had recently presented in the A-Edition Introduction to the first critique. Indeed, the *Metaphysik Mongrovius*'s characterization of the new critical metaphysics as a science of the a priori principles [*principiorum*] of human cognition, together with its characterization of a *principium*, sheds much light on Kant's characterization of the project of the first critique as the self-tribunal of pure reason.

3. The Necessity, Generality, and Certainty of A Priori Cognition

I want now to turn to texts of the first critique, beginning with Kant's characterizations of the a priori in the Introductions. The present section

[84] Ibid. [85] 29: 752.

aims to dispel the impression that in the Introductions Kant defines the term 'a priori,' in the sense in which he will employ it, simply in terms of independence from experience, and to provide some initial motivation for my claim that his core notion of the a priori is the from-grounds one. It must be admitted at the outset that neither the A- nor the B-Edition Introduction provides an explicit from-grounds characterization of the a priori. However, as we will see, the characterizations Kant offers in terms of generality, necessity, and certainty are standard in Leibnizian view of the a priori; indeed, when we attend to these characterizations carefully, we can see that they portray a priori cognition in terms of the apodictic consciousness of the necessity with which consequences follow from determining grounds, a consciousness had in and through rationally perceiving *rationes essendi* as such. In short, these characterizations of a priori cognition tacitly express the from-grounds notion standard in the Leibnizian tradition. Especially in connection with the B-Edition Introduction, I will also be concerned, to some extent, with the positive characterization of the a priori he offers in genetic terms, but will defer detailed discussion of this characterization's relation to the from-grounds notion to the next section.

Consider how Kant first introduces the term 'a priori' in the Introductions to the first critique. In the first paragraph of the A-Edition Introduction, Kant flatly rejects the empiricist's restriction of our understanding to the field of experience. Echoing Leibniz's contention in the *New Essays* that experience can, through induction, only yield truths that we presume to be necessary, Kant writes

> it [experience] is far from the only field to which our understanding can be restricted. It tells us, to be sure, what is, but never that it must necessarily be thus and not otherwise. For that very reason, it gives us no true generality, and reason, which is so desirous of this kind of cognitions, is more stimulated than satisfied with it. Now such general cognitions, which at the same time have the character of inner necessity, must be clear and certain for themselves, independently of experience; hence one calls them a priori cognitions; whereas that which is merely borrowed from experience is, as it is put, cognized only a posteriori, or empirically.[86]

And in the B-Edition Introduction, having in the first two paragraphs distinguished the claim that all our cognition 'commences [*anhebt*] with

[86] A1–2.

experience' from the empiricists' claim all our cognition 'arises [*entspringt sie aus*] from experience', Kant continues:

> It is therefore at least a question requiring closer investigation, and one not to be dismissed at first glance, whether there is any such cognition independent of all experience and even of all impressions of the senses. One calls such cognitions a priori, and distinguishes them from empirical ones, which have their sources a posteriori, namely, in experience.

Moreover, Kant goes on to distinguish between cognitions that derive immediately from experience and those that derive immediately from a general rule which we have borrowed from experience. The latter, unlike the former, are independent of some particular experiences. But such cognitions are, nonetheless, not absolutely independent of experience. He then remarks, 'In the sequel therefore we will understand by a priori cognitions not those that occur independently of this or that experience, but rather those that occur *absolutely* independently of all experience.'[87] Kant is often read in these passages as defining the term 'a priori', in the sense in which he will use it, simply in terms of independence from experience, and 'a posteriori', correlatively, in terms of dependence on experience. The passages from the B-Edition Introduction, especially, do appear on first pass to advance such a definition. They no doubt played an important role in changing the standard philosophical usage of the terms 'a priori' and 'a posteriori'—namely, in supplanting the from-grounds and from-consequences senses of these terms with those simply in terms of independence and dependence on experience. And it is, I think, primarily the passages at B2 that have lead many commentators to read Kant as, in his own usage, making this clean break from the old senses of these terms.

However, closer examination will show that these passages not only need not, but should not, be read as *defining* the senses in which Kant will use the terms 'a priori' and 'a posteriori' simply in terms of a representation's relation to experience.

Consider first how Kant introduces the term 'a priori cognition' at A2. Independence from experience does figure in Kant's characterization here of the cognitions due to which (note the 'hence') 'one calls them a priori cognitions'. So Kant here seems to be reporting as standard a nominal

[87] B2.

definition (*Wortbestimmung*) of the a priori in which independence from experience plays a part. But so do necessity, generality, and certainty. Indeed, the independence from experience first invoked in this passage makes up part of Kant's characterization of the certainty of a priori cognition, where the certainty in question is evidently a normative notion, one that concerns the nature of the cognition's justification. So at A2 Kant does not privilege a characterization simply in terms of independence from experience over other characterizations as what itself suffices to define what is commonly meant by 'a priori'.

We need, then, to attend to the entirety of the initial characterization of a priori cognitions that at A2 Kant says forms the basis on which one calls them a priori cognitions. This characterization concerns the intrinsic clarity and certainty enjoyed by these cognitions: in Kant's words, they are 'clear and certain for themselves, independently of experience'. Note the implication that these cognitions are clear and certain independently of experience *because* they are clear and certain *for themselves*; moreover, Kant seems to infer this independence from experience in the service of bringing these cognitions' intrinsic clarity and certainty into clearer view. The question, then, is, what according to Kant was commonly understood in the relevant circles by this intrinsic clarity and certainty? The answer, I propose, is the clarity and certainty that a cognition enjoys as a from-grounds cognition, that is, as an ontological, determining ground that is registered as such in a cognitive consciousness. The intrinsic clarity and certainty of a cognition is, in a word, apodictic.

That this is the correct understanding of the intrinsic clarity and certainty Kant invokes becomes evident when we attend to how he derives this inner clarity and certainty of a priori cognitions from his earlier characterizations of these cognitions in terms of necessity and true generality. Kant introduces these cognitions as ones that 'teach us that something must necessarily be thus' and 'therefore have true generality', that is, as 'general cognitions, which at the same time have the character of inner necessity'. It is not, on first pass, entirely clear what Kant means by this 'inner necessity'. But our examination of the notion of the a priori that enjoyed common currency in the Leibnizian tradition strongly suggests the following: an a priori cognition consists in the consciousness of the distinctive character of a *ratio essendi*, and so of the necessity with which it determines its consequence; an a priori cognition thus is a necessity that is, as such, and in its distinctive

character, registered in a cognitive consciousness, whereas an a posteriori cognition posits a *ratio essendi* whose necessity is not itself registered in that cognition. Notice that this positive characterization of a priori cognition in terms of necessity derives from the from-grounds conception of the a priori. If this is what Kant has in mind by a general cognition's 'having the character of inner necessity', we can see how at A2 he is drawing on this characterization to specify the generality distinctive of a priori cognition: a generality that is grasped, in that cognition, in respect of how it follows from a *ratio essendi*: after all, whatever is essential to a thing of a certain kind is had necessarily by all instances of that kind. Note that on this reading Kant does not, in this passage, specify the sort of generality characteristic of a priori cognition only negatively, by distinguishing it from the generality which derives merely from induction over individual cases given in experience. Moreover, on this reading, we can see how Kant understands the intrinsic clarity and certainty of a priori cognitions in such a way that it follows from their having 'the character of inner necessity': it is because these cognitions consist in a consciousness of *rationes essendi* in respect of their character as antecedently determining grounds, and so of how they necessitate their consequences, that they 'must be clear and certain for themselves, independently of experience'. Our a posteriori cognitions, in contrast, have what clarity and certainty they enjoy only on the assumption that the regularities among appearance they represent are consequences of *rationes essendi* that they do not register in their distinctive character, and in such a way that these appearances constitute experience.

That Kant does not, at the opening of the Introductions, mean simply to equate the a priori and the non-empirical is also consistent with the way, in the above-cited passages, he contrasts the cognitions he says are 'called a priori cognitions' with those cognitions that we derive from experience. In particular, 'cognized only a posteriori, or empirically' at A2 does not have to be read as making 'empirical' appositive to 'a posteriori'. Kant could mean 'or empirically' merely to specify the respect in which this cognition is a posteriori, in the from-consequences sense: what is 'merely borrowed from experience' is 'cognized only a posteriori' in that the consequences that are registered as justifying grounds in this cognition are the effects things have on our sensibility in virtue of which we experience them. Moreover, when at B2 Kant contrasts the cognitions he says are 'called a priori' with empirical cognitions, he uses 'a posteriori' in a way that not

only need not be, but does not seem to be, synonymous with 'empirical'. When he characterizes empirical cognitions as those which 'have their sources a posteriori, namely in experience', the force of the phrase 'namely in experience' seems to be to specify one sort of a posteriori source. Kant is, I think, here naturally read as making the same point he makes in the passage from the *Metaphysik Mongrovius* discussed in the previous section—namely, that empirical cognitions are a posteriori inasmuch as the consequences from which these cognitions proceed are the effects of the objects of these cognitions on our sensibility in virtue of which they are given in experience.

But, the reader might well ask, what about Kant's opening characterization of a priori cognition in the B-Edition Introduction? Doesn't Kant at B2 repeatedly and flatly define the term 'a priori' simply as 'non-empirical'? One might think that, if I am right about the A-Edition Introduction, perhaps what I have uncovered is not a commitment on Kant's part to the from-grounds notion of the a priori that persists throughout the critical period, but rather a sharp shift in substance between the sense of 'a priori' that Kant specifies in the A-Edition, and the sense he specifies in the B-Edition.

At this point, we need to attend carefully to what, at the opening of the above-cited passage from B2, Kant is referring back to as 'such cognition independent of all experience and even of all impressions of the senses'. Here is what immediately precedes the passage:

> Although all our cognition commences with experience, yet it does not on that account all arise from experience. For it could well be that even our experiential cognition [*Erfahrungserkenntnis*] is a composite of that which we receive through impressions and that which our own cognitive capacity (merely prompted by sensible impressions) provides out of itself, which addition we cannot distinguish from that fundamental material until long practice has made us attentive to it and skilled in separating it out.[88]

'Such cognitions' later in B2, thus, refers back to cognitions 'which our own cognitive capacity provides out of itself'. Two important points follow. First, unlike in the A-Edition, Kant opens the B-Edition Introduction with a characterization of a priori cognition that is, explicitly at least, one

[88] B1–2.

simply in terms of its genesis: what is in question is a priori cognition's having a genesis that is independent from 'all experience and even of all impressions of the senses'. Second, Kant's initial characterization of this origin is positive: the subsequent characterization of a priori cognitions in terms of experience is clearly intended to supplement and clarify Kant's initial, and primary, characterization of these cognitions simply as ones that 'our cognitive capacity provides out of itself'. Thus, even if Kant is, at B2, offering a definition of the term 'a priori' different from that which he offers us at A2, he is not offering the one that is familiar to us—namely, one simply in terms of having a justification that is independent of experience.

However, on closer examination, it seems likely that Kant does not at B2 mean to be offering a nominal definition of a priori cognition at all. Note that Kant does not, as he does at A2, say that it is due to the characterization of a priori cognition he has provided that 'one calls' these cognitions a priori: omitting the 'hence' (*daher*) we find at A2, he writes simply that 'one calls such cognitions a priori', where the relevant cognitions are described as ones that 'our own cognitive capacity provides out of itself', and so cognitions 'independent of all experience and even of all impressions of the senses'. Now, as we will see in the next section, Leibniz and his successors would all call cognitions that fit this description 'a priori'. For they all held, as a substantive thesis, that only cognitions that are inherent in our cognitive capacity, and thus cognitions that we do not derive from experience, could enjoy the apodictic clarity and certainty characteristic of the a priori, in the from-grounds sense of 'a priori'. And, as we will also see in detail in Section 4, this is a thesis that Kant advances in the opening paragraphs of both Introductions. But Leibniz and his successors would not have offered 'arising simply from one's capacity of cognition' as the nominal definition of 'a priori'. It seems plausible, then, to assume that Kant, at B2 as at A2, is making only the claim that, within the Leibnizian tradition, cognitions with this origin would have been standardly referred to as a priori cognitions. He should not be read as here offering a nominal definition of 'a priori' at all, let alone one simply in terms of independence from experience.

What, then, accounts for the difference between the ways in which Kant introduces a priori cognition in the A- and B-Editions? I suggest that Kant chooses to introduce a priori cognitions in the B-Edition as cognitions that one's capacity of cognition 'provides out of itself', in part because

he wants to focus attention on the positive characterization of a priori cognition which, when correctly specified, will prove in the sequel to be his real definition of a priori cognition, the definition that specifies what makes an a priori cognition an a priori cognition and in terms of which he explains the real possibility of our a priori cognition.[89] This shift brings the Introduction more in line with Kant's own understanding of proper philosophical methodology: philosophy ought not to be concerned with mere nominal definitions (*Wortbestimmungen*), which as such are merely arbitrary, but rather ought to explicate concepts that are given a priori by the very nature of our capacity of cognition with the aim of arriving at a complete exposition (*Exposition*) of those concepts.[90] Moreover, initially characterizing a priori cognition in terms of its origin allows Kant, in the B-Edition Introduction, to broach the issue of a priori cognition first in direct connection with his pivotal contention that our experience consists, in part, of a content that does not itself derive from experience, but rather from our own cognitive capacity—indeed, that this content constitutes a form common to any possible experience in general and which conditions the possibility of any experience as its *ratio essendi*. In the A-Edition, Kant introduces this thought only after having introduced the features of the a priori, in the from-grounds conception of the a priori, that are salient to the common rationalist claim that experience cannot yield a priori cognition. Indeed, it seems likely that Kant revised the Introduction for the B-Edition in part so that it would more squarely engage those in the empiricist tradition. We have seen that the A-Edition Introduction's opening paragraph presupposes that our understanding is capable of a priori cognition understood in the standard Leibnizian sense. But the empiricists would have denied that we are capable of cognitions with the features

[89] I here understand 'definition' in the technical sense of '*Erklärung*' that Kant specifies at A727/B755–6, on which conditions of *Erklärungen* are less exacting than those on *Definitionen*; Kant holds that we can provide *Definitionen* only in mathematics; in doing philosophy, we can only provide *Erklärungen* (A729/B757). Also, I am not suggesting that Kant at B1 presents his positive genetic characterization of the a priori as a real definition. Indeed, doing so would not be in line with Kant's own account of the methodology appropriate to philosophy. In this account, philosophy, unlike mathematics, ought not begin with real definitions (A730/B758–9). It is only after having analyzed given a priori concepts, and succeeded in achieving an entirely a priori insight through them, that the philosopher is in the position to offer any characterizations as adequate to her subject-matter, and so as real definitions. For, in philosophy, it is only such an insight that reveals the real possibility of the subject-matter in drawing on the analyzed concepts to provide a complete a priori explanation of this possibility.

[90] A728/B756.

of necessity, generality, and certainty Leibnizians specify.[91] The B-Edition Introduction opens, instead, with a genetic characterization of the a priori, because it opens by raising a challenge for empiricists: perhaps experience itself is 'a composite of that which we receive through the sense and that which our own cognitive capacity (merely prompted by sensible impressions) provides out of itself'.[92]

Kant, of course, also offers his characterizations of the a priori in terms of necessity and true, or strict, generality in Section II of the B-Edition Introduction, in the course of arguing that we possess a priori cognition, and indeed that 'even the common understanding is never without them'.[93] He simply postpones introducing these characterizations until after he has, in Section I, introduced a priori cognitions, and distinguished pure cognitions from empirical ones—all in terms of the epistemically salient origins of these cognitions. Attending carefully to these characterizations of the a priori in the opening paragraph of Section II will confirm both my contention that these characterizations present salient features of the from-grounds conception of the a priori and my suggestion that Kant's characterizations of the a priori in the B-Edition focus on bringing into clear relief what will serve as his real definition (*Erklärung*) of a priori cognition.

What we need to understand is what, exactly, Kant means when he specifies necessity and strict generality as 'secure indications [*Kennzeichnen*] of an a priori cognition'.[94] In particular, what Kant points to as such indications are a proposition's being thought 'along with its necessity'—that

[91] Indeed, in the B-Edition, Kant first mentions the necessity and strict generality of an a priori cognition, as its 'secure indications', only in section II. He does so immediately after having introduced, at the end of section I, a demanding genetic sense of 'a priori' that he terms 'absolute'—one in which only cognitions that 'occur independently of any experience' count as a priori—and having specified that this is the sense in which he will understand the term in what follows (B2). Moreover, in specifying this sense, he does not presuppose that we have, or even could have, any a priori cognition in this sense: the heading of section II is 'We are in possession of certain a priori cognitions'. All this confirms that Kant intends the B-Edition Introduction to engage empiricists, such as Hume, who would have denied that we have a priori cognition where 'a priori' is understood in the absolute sense. This suggests that in writing 'in the sequel therefore I will understand by a priori cognitions not those that occur independently of this or that experience, but rather those that occur absolutely independent of all experience' (B2) he is not offering a stipulative definition of the term, as he is commonly taken to be. Rather, he is specifying a demanding understanding of the a priori on which the claim that we are capable of a priori cognition would be hotly contested by empiricists. Notice, further, against the standard reading of this sentence as offering Kant's non-empirical definition of the a priori, that it is only in specifying the absolute sense of 'a priori', and not in specifying the most general sense of 'a priori', that he here invokes independence from experience.

[92] B1.

[93] B3.

[94] B4.

is, with the thought of its distinctive necessity, and not just the thought of necessity in general—and its being 'thought in strict generality'—that is, as having a generality that allows for no exceptions whatsoever to be possible (and not simply in allowing for no exceptions in the cases that have been observed so far).[95] Moreover, Kant claims that whereas the generality of a judgment that is derived from induction results from 'an arbitrary increase' in the validity of that judgment, strict generality 'belongs to a judgment essentially'.[96] Kant's thought, then, seems to be that the validity of judgments that are strictly general differs in kind from the validity of judgments that enjoy merely assumed and comparative generality: the former, unlike, latter, consists in a judgment's appealing, as such, to grounds that suffice to establish strict generality. He then claims that the essentiality of a judgment's strict generality 'points to a special source of cognition for it, namely a capacity of a priori cognition' and infers from this claim that the two properties are 'secure indications of an a priori cognition and also belong together inseparably'.[97]

These claims and inferences echo those that we saw Leibniz make in distinguishing propositions of fact and propositions of reason.[98] Moreover, as we saw in connection with Leibniz, these claims and inferences implicate the from-ground conception of the a priori: for they make sense, on the hypothesis that Kant is here taking our cognizing something a priori to be equivalent to cognizing it out of a rational perception of the necessity with which it follows from its ontological determining ground. To see how certain predicates belong to, or follow from, some determinate essence and so from the mere possibility of a determinate kind is to see the necessity and strict university with which they apply to all possible instances of that kind. In short, what Kant takes to be characteristic of our a priori cognition is its consisting of the consciousness, as such, of antecedently determining grounds, and so of the necessity with which these grounds serve to determine the truth of some particular strictly general judgment. Moreover, Kant's calling these characterizations 'secure indications' is plausibly taken to be a warning against regarding them as what, severally or jointly, constitute the real definition of a priori cognition. Kant is claiming that

[95] B3. [96] B4. [97] B4.

[98] NE 446; cf. Section 1, above. Burge makes this point at 'Frege on Apriority', 23 n., in the course of an insightful discussion of how necessity and strict generality are, for Kant, secure indications of a priori cognition.

they are, in Leibniz's terminology, merely reciprocal concepts that as such are sufficient to identify something.[99] And in saying that these 'secure indications' point to 'a special source of cognition'—namely, 'a capacity of a priori cognition',[100] Kant is hinting at the distinctive origin of a priori cognition in terms of which he will, in the Transcendental Analytic, explain a priori the possibility of our having cognition of objects that is a priori in the from-grounds sense of 'a priori'. This explanation will take the form of a transcendental deduction—that is, an account of the origin of our a priori concepts of objects in an exercise of our capacity of a priori cognition that serves to establish their a priori objective validity in respect of possible objects of our experience.

Close attention to Kant's discussions in the first critique's Transcendental Doctrine of Method of the certainty characteristic of a priori cognition also confirms that the from-ground conception of the a priori underlies this characterization. Consider first how Kant's discussion of direct, or ostensive, proof in the Transcendental Doctrine of Method confirms and develops his implication at A2 that what characterizes the a priori is not certainty per se, but a certain sort of certainty—namely, one that rests on a grasp of how *rationes essendi* exclude certain possibilities, and ground necessary truths.

> The direct or ostensive proof is, in all kinds of cognition, that which is combined with the conviction of truth and simultaneously with insight into its sources [*Quellen*]; the apogogic proof, on the contrary, can produce certainty, to be sure, but never comprehensibility of the truth in regard to its connection with the grounds of its possibility. Hence the latter are more of an emergency aid than a procedure which satisfies all the aims of reason.[101]

Here Kant is explicit that the sort of certainty, and clarity, achieved in direct proof differs from other sorts of certainty and clarity. And he characterizes this certainty in terms of conviction born of insight into, that is, of cognizing

[99] I agree, then, with commentators who warn against reading Kant, in section II of the B-Edition Introductions, as *defining* the a priori in terms of necessity and true or strict generality (for a particularly careful and nuanced discussion of this point, see ibid., 23). But, as I have been suggesting, we need at this juncture to distinguish nominal and real definitions, and recognize that Kant's concern here is with real definitions, and not with nominal ones (*Wortbestimmungen*). Kant's project in the first critique requires him to develop a conception of our a priori cognition that can serve as a real definition, that is, one that allows us to exhibit a priori, or from its ontological grounds, the real possibility of our a priori cognition.

[100] B4.

[101] A789–90/B817–18.

a priori in the from-grounds sense, the sources (read: principles) of what is proved. Moreover, Kant explicates this insight as comprehension of how the possibility of the truth that is proved necessitates this truth. This is to make *rationes essendi* the sole grounds that we can, in a direct proof, rationally perceive as such, so as to have insight into the source of a truth. It is, moreover, to make consciousness of necessitating grounds, as such, an identifying mark of the a priori: for a truth that follows from the grounds of its possibility is a necessary truth. Now earlier, at A755/B803, Kant has connected proving a priori and apodictic certainty—echoing Leibniz's equating a priori proof with *apodeixis*.[102] The certainty characteristic of direct proof, then, is apodictic certainty.[103] And Kant's remark here that only direct proof satisfies the aims of reason expands on his remark at A2 that reason is only satisfied with a priori cognition.

For our purposes, it will be important to clarify and develop the link Kant draws in these passages between reason and the a priori. This link is, in the first instance, one between a priori proof and cognizing through reason.[104] To cognize through reason, in the relevant sense, is not simply to use one's reason in cognizing. It is, rather, to cognize in and through a rational perception of ontological determining ground as such. Kant's link between reason and the a priori thus parallels that drawn by Leibniz when, as we saw above, the latter equates a priori truths with truths of reason. In order to bring out this point, and to clarify his notion of pure reason, I want now to attend to Kant's characterizations of reason as a capacity of principles in light of the characterization of a principle he provides in the *Metaphysik Mongrovius*. What will come to light is how Kant conceives of reason, in its real theoretical use, as the capacity to appreciate ontological determining grounds as such and so in respect of how they determine some particular consequence.

In the opening of the Transcendental Dialectic, Kant proposes to characterize reason, in a sense that contrasts reason with the understanding, by appeal to the notion of a principle (*principium*) (A299/B356).[105] In the course of doing so, he distinguishes several successively more demanding senses of 'principle', finally arriving at the absolute sense required by his

[102] C 408.

[103] In the Jäsche *Logic*, he also distinguishes rational, as against empirical, certainty 'by the consciousness of necessity connected with' rational certainty, and then goes on to remark that rational certainty is 'thus an apodictic certainty' (9: 71).

[104] This link also comes to expression when, in the Jäsche *Logic*, Kant equates cognizing a priori, or having insight, with cognizing through reason (9: 64).

[105] A299/B356.

characterization of reason. Most relevant to our purposes is the distinction Kant draws between the least and all the more demanding senses, a distinction that illumines his notion of cognizing absolutely a priori. In the least demanding sense, any general cognition, be it a posteriori or a priori, counts as a principle, simply in virtue of the fact that our reason can employ it as a major premise in a syllogistic inference.[106] For in doing so I regard a general a posteriori cognition, such as 'Bodies are heavy' as the ground of a cognition of some particular thing as heavy constituting a ground that determines that thing's being heavy as a consequence. I thereby assume that this general cognition a posteriori, itself a mere teaching of experience, is strictly general and necessary. Kant then specifies a more demanding sense of 'principle' on which a general cognition must itself be a principle, and claims that a principle in this sense must be a general cognition a priori.[107] He gives as examples of such principles the principles of mathematics and the principles of pure understanding. An a priori general cognition is itself a principle, because it is not simply in a use, one in which I merely assume its strict generality, that it grounds a ground that follows a certain rule. Rather, such a general cognition is itself a cognition of this strict generality, a rational perception of how *rationes essendi*, as ontological grounds of possibility, necessitate some determinate consequence.

In distinguishing these senses of 'principle', and the corresponding senses in which we can be said to cognize through reason, Kant is in effect distinguishing merely relative and absolute senses of 'cognizing a priori', using the from-grounds notion of the a priori. Recall how, in the B-Edition Introduction, Kant draws the distinction between merely relative and absolute senses of 'cognizing a priori' in terms of independence from experience. Kant remarks that one can be said to know a priori that one's house, once its foundations have been undermined, will collapse, when one knows this prior to experiencing its collapse.[108] But, if one knows this only on the basis of one's past experience of unsupported structures eventually collapsing, and so from reasoning according to a general rule, derived by induction from this experience, to the effect that unsupported structures all eventually collapse, one does not know the proposition in question absolutely a priori. Absolutely a priori knowledge that one's house, once its foundations have been undermined, will collapse would require that one

[106] Ibid. [107] Ibid. [108] B2.

know this absolutely independently of any experience.[109] This is the point that prompts Kant to specify that 'In the sequel therefore we will understand by a priori cognitions not those that occur independently of this or that experience, but rather those that occur absolutely independently of all experience'.[110] Kant's discussion of principles indicates that this distinction between two senses of 'cognizing a priori', one absolute and the other merely relative, is more fundamentally one between cognizing through an a priori principle and cognizing through an a posteriori principle—between cognizing that is, and cognizing that is not, founded on a rational perception of an ontological ground as such. When one employs a general cognition a posteriori as a principle, one infers a particular instance of a consequence from a particular instance of a ground, and so cognizes that particular consequence from its ontological ground. So, in a sense, one does cognize something a priori, and cognizes it through one's reason. But one cognizes this consequence a priori only relative to this particular ground: absolutely speaking, one cognizes this consequence a posteriori, in so far as one cognizes it through an a posteriori principle, and so from the consequences that one employs as a ground of cognition in cognizing this principle. Cognizing something absolutely a priori requires an a priori principle, and so insight into an ontological ground. In Kant's example, it would require rational perception of the way in which the undermining of a structure necessitates its collapsing. And this suggests that when Kant remarks that 'we will understand by a priori cognitions . . . those that occur absolutely independently of all experience',[111] he is simply specifying the absolute sense of apriority in terms of an absolute independence from experience; he need not, and I think should not, be read as offering a definition, real or nominal, of a priori cognition in terms of independence from experience.

We can now see that Kant retains a rationalist conception of science, in the proper sense, as demonstrative, a conception that connects apodictic certainty, reasoning according to a priori principles, and consciousness of necessity. In this conception, a genuine science (*Wissenschaft*) is a body of cognitions that reason unifies into a system, under a priori principles and so in a systematic body of absolutely a priori cognition:

> What can be called proper science [*Wissenschaft*] is only that whose certainty is apodictic; cognition that can contain mere empirical certainty is only knowledge

[109] B2. [110] Ibid. [111] Ibid.

[*Wissen*] improperly so called. Any whole of cognition that is systematic can, for this reason, already be called science, and, if the connection of cognition in this system is an interconnection of grounds and consequences, even rational science. If, however, the grounds or principles themselves are still in the end merely empirical, as in chemistry, for example, and the laws from which the given facts are explained through reason are mere laws of experience, then they carry with them no consciousness of their necessity (they are not apodictically certain), and thus the whole of cognition does not deserve the name of science in the strict sense.[112]

Note that science proper, in virtue of enjoying apodictic certainty, is what constitutes human knowledge (*Wissen*) in the proper sense. Note, too, that Kant links consciousness of the necessity of laws—something lacking in the case of mere laws of experience—with apodictic certainty. Laws of experience are, as laws, necessary rules. But we cannot have insight into the necessity of a particular empirical law. This provides further confirmation that what is distinctive of apodictic certainty, and so of a priori cognition, is the consciousness of *rationes essendi* in respect of how they necessitate certain truths.

We have seen that Kant does not put forward just any necessity, generality, or certainty as an identifying characteristic of the a priori. The 'secure indications' of the a priori are not simply the necessity and generality that are presumed in a posteriori principles, but rather necessity and university that are, or can be, rationally perceived as such in an act of insight. They are, in other words, the necessity and generality that, according to the Leibnizian tradition, characterize the a priori cognition of a finite mind, because the only ontological grounds that a finite mind can cognize are *rationes essendi*. Moreover, the certainty distinctive of our a priori cognition is not just any certainty, but an apodictic and rational certainty, one that derives from the rational perception of ontological grounds as such—*apodeixis*. In short, the necessity, generality, and certainty that Kant takes to be characteristic of our a priori cognition all derive from an account of the a priori cognition of a finite mind that employs the now archaic, from-grounds notion of the a priori.

But why, then, doesn't Kant explicitly articulate the from-grounds notion of the a priori in the Introductions to the first critique? A possible explanation is that stating it would have been misleading, given the

[112] MFNS; 5, 468.

distinctive character of the new critical approach to metaphysics he is presenting in the Introductions. This approach begins by raising, as a problem, our reason's claim, in pursuing the sciences of logic, metaphysics, mathematics, and natural philosophy, to generate genuine synthetic a priori cognition. One account of what reason is properly regarded as thereby laying claim to is genuine insight into things in themselves—ontological grounds in an order of being that is absolutely independent of our cognition. Indeed, this is the common account within the Leibnizian tradition, and just the account that Kant wants to deny. The revolutionary claim of Kant's critical account of our a priori theoretical cognition is that, contrary to traditional rational metaphysics, the a priori principles of our cognition are not cognitions of things, when we consider things in respect of a reality they have that transcends possible human experience. They are, rather, cognitions that merely as representations—representations that constitute the form of our experience—make appearances grounds of cognition of things as they appear. Specifying the from-grounds notion of the a priori, then, would have given Kant's audience the misleading impression that he was, following the Leibnizian tradition, building into his initial, working notion of the a priori just the absolutely a priori cognition of things in themselves that the critical philosophy denies us.[113]

4. The Genesis of A Priori Cognition and the Spontaneity of Cognition

I want now to attend more closely to Kant's genetic characterization of a priori cognition, on which a priori cognition is cognition that derives, not from sensory experience, but rather from the exercise of our capacity of cognition itself. Both the sense and the motivation of this characterization

[113] But why, given the prominence of the from-grounds notion of the a priori in his time, didn't Kant explicitly specify that he was working with this notion (perhaps deriving the non-empirical characterization of the a priori from it) and then explain how he wasn't committed to theoretical cognition of things in themselves? (Thanks to Larry Jorgensen for this question.) A possible answer is that explaining how he wasn't committed to theoretical cognition of things in themselves would have required introducing the transcendental sense of the distinction between things as they appear and things in themselves, and Kant, quite naturally, wanted to hold off introducing this distinction until later in the work. After all, it isn't until later in the Introductions that Kant even introduces the notion of the transcendental (A11/B25).

will be illumined by the recognition that Kant operates with the from-grounds notion of the a priori. Moreover, it will, once again, prove useful to have Leibniz's account of the a priori in mind as we examine Kant. In holding that a priori cognition, understood in the from-grounds sense, must derive from the capacity of cognition itself, Kant follows Leibniz, although, to be sure, Leibniz's and Kant's accounts of the nature of our cognitive capacity and of the manner in which a priori cognition derives from it differ in important ways. Attending closely to Kant's account of the genesis of our a priori cognition—on which all our a priori cognition, even our purely a priori sensible cognition, must originate in an exercise of our understanding, the spontaneity of cognition—will put us in a position to see how the genetic characterization of the a priori is an a priori cognition of a priori cognition. Indeed, it is the real definition that Kant employs, over the course of the Analytic, to exhibit a priori the real possibility of pure reason, and thereby of all our a priori cognition.

Before turning to Kant's genetic characterization of a priori cognition, it will prove useful to set the stage with a sketch of Leibniz's account of the genesis of a priori cognition. Leibniz lays out this account in the *New Essays*, so it was one with which Kant and his intended audience were well acquainted. I will follow this sketch with a brief, initial statement of the crucial respect in which Kant's own account of this genesis departs from Leibniz's.

We have already seen that, in Leibniz's view, a priori cognition must enjoy an origin other than sensory experience: our cognition of a priori truths is cognition that reflection alone provides, independently of sense. Leibniz holds, further, that this cognition—to which he refers as 'thoughts of reflection'—originates in the mind's reflective awareness of its own nature: 'The mind must at least give itself its thoughts of reflection, since it is the mind which reflects'.[114] This cognition includes, crucially, certain concepts of general metaphysics, such as substance and unity:

> Reflection is nothing but attention to what is within us, and the senses do not give us what we carry with us already. In view of this, can it be denied that there is a great deal that is innate in our minds, since we are innate to ourselves, so to speak, and since we include Being, Unity, Substance, Duration, Change, Action, Perception, Pleasure, and hosts of other objects of our intellectual ideas? And since

[114] NE 119.

these objects are immediately related to our understanding and always present to it (although our distractions and needs prevent us being aware of them), is it any wonder that we say that these ideas, along with what depends on them, are innate in us?[115]

What are innate in us, in Leibniz's view, are not actual thoughts, but rather ideas, which he characterizes as 'dispositions' or 'potentialities' for thought.[116] Moreover, Leibniz holds, with Plato, that all ideas are innate in us.[117] We can be said to receive cognition from external things through our senses only in so far as the reasons that determine our souls to form certain thoughts are expressed more clearly in those things than in ourselves.[118] Indeed, in Leibniz's monadology, sense is fundamentally nothing but that which renders perceptions obscure and indistinct, the primitive passive potency of perception.[119] What distinguishes the intellectual ideas from the ideas of sense, then, is that they are dispositions for thought simply of our understanding and which thus determine the content of our thoughts of reflection independently of our primitive passive potency for perception.[120] This suggests that when at NE 51 Leibniz claims that reflection is 'nothing but attention to what is within us', and that 'the senses do not give us what we carry with us already', he means by 'what is within us' what is within us only in so far as we constitute minds, the subjects of understanding.[121]

Consider next that, in Leibniz's account, what distinguishes minds, or subjects of understanding, from other monads is their ability to reflect, that is, through the representation 'I' to be immediately aware of what one is and what one does.[122] Moreover, he maintains that a mind's ability to reflect is what allows it to discover necessary and universal truths.[123] This

[115] 51–2; cf. *Monadology*, § 30. [116] NE 52, 86; cf. DM § 26. [117] DM § 26.
[118] DM § 27. [119] *Leibniz*, chap. 13, esp. sect. 2.
[120] Notice the consequence that thoughts of reflection are clear and distinct.
[121] This is confirmed by Leibniz's going on, in the next sentence at NE 51, to equate 'what is within us' with 'what is innate in our minds', and by his remark at NE 119 that the *mind* must give itself its thoughts of reflection. In taking 'in us' to have a restrictive sense in these passages, I am suggesting a tack similar to Mark Kulstad's: Kulstad proposes that at NE 51 Leibniz speaks of what is 'in us' in the 'image-excluding' sense he employs at NE 52–3 and 76; see ch. 4 of his *Leibniz on Apperception, Consciousness, and Reflection* (Munich: Philosophia, 1991), esp. 123–6. But on my reading the relevant sense of 'in us' is 'in us insofar as we are subjects of understanding'. In favor of my reading, note that at NE 76, the contrast between understanding and sense seems to be operative: Leibniz excludes from what is 'in us' not all images, but 'sensible images'. And, at NE 52–3, Leibniz contrasts what is in our soul with 'the images that are borrowed from outside it'—that is, images that are the result of the exercise of our primitive passive potency. [122] DM § 34.
[123] Ibid.

makes it clear enough that Leibniz regards thoughts of reflection as a priori cognition. But, to see why he does, we need to appreciate how, in Leibniz's account, thoughts of reflection are a priori cognitions in the from-grounds sense. And here the key point to see is that, in this account, the exercise of the intellectual ideas in a reflective consciousness of its own thinking is essential to a mind's enduring throughout a succession of perceptions and desires as the individual substance that is their common subject. This, I propose, is why he holds that the objects of our intellectual ideas—Being, Unity, Substance, Duration, Change, Action, Perception, Pleasure—are 'related immediately to the understanding and always present to it'. In his Correspondence with Arnauld, Leibniz claims that there must be a reason a priori that makes it the case that 'I, who was in Paris, am now in Germany', and that this reason can only be 'what is called I, which is the foundation of the connection of all my different states'.[124] What the thoughts of reflection represent immediately, albeit only in general and in respect of what is common to all possible minds, is how the exercise of this a priori ground constitutes it as a persisting individual substance. And this just is to say that these thoughts are a priori cognitions, in the from-grounds sense of 'a priori'. When we draw on reflection to become clearly and distinctly aware of how Being, Unity, Substance, and so forth are essential to us, we come to know certain necessary and universal truths of metaphysics. Notice that, in Leibniz's account, the origin of thoughts of reflection in the activity that constitutes the numerical identity of the thinking subject is essential to their constituting a priori cognition. Notice, too, that one's purely a priori insight, through thoughts of reflection, reaches only to the *rationes essendi* of any mind, in general. It does not, in particular, reach to the *rationes fiendi* of one's own existence as the distinctive individual contingent being that one is.[125]

[124] AG 73.

[125] This paragraph develops Adams's important observation that the sense of 'a priori' in play at AG 73 must be the from-grounds one (*Leibniz*, 77). To see why thoughts of reflection do not reach to what distinguishes one as an individual substance, one needs to keep in mind that, for Leibniz, a created monad's primitive passive potency is, as a principle of intraspecific individuation, what determines its numerical identity. Here I draw on Adams's interpretation of primary matter in *Leibniz*, ch. 12, esp. section 1.3. Notice that the reading I propose explains why Leibniz holds, not just that there must be an a priori ground of the identity of the I, but that this ground must be given to us, as such a ground, in a priori cognition. It is in virtue of the way in which a person's identity is in this way always 'apparent' to that person that he constitutes a person (NE 236). This last point, however, must be understood in such a way as to allow one to remain the same person through a 'gap in recollection' (ibid.).

Kant challenges Leibniz's claim that what reflection yields is cognition through which we can achieve genuine purely intellectual insight into our nature as a thinking thing. But he agrees with Leibniz that reflection alone does yield purely intellectual a priori cognition, and in the from-grounds sense of 'a priori'. Where he parts company with Leibniz is in insisting that this cognition is, of itself, a priori cognition only of a priori cognition: that is, through this cognition—in particular, the pure concepts of the understanding—we can achieve genuine purely intellectual insight only into the ontological grounds of the possibility of our a priori cognition, not into the ontological grounds of the possibility of a thing in itself. Leibniz, and other practitioners of rational psychology, failed to see that the thoughts of reflection can constitute a priori cognition of objects, as against cognition merely of the subject as such, only in relation to possible sensory experience. Of themselves, concepts such as *unity* and *substance* constitute mere thought, and not genuine cognition, of objects in general. These concepts come to constitute cognitions of an objective unity of representations, one that characterizes things in respect of their objective reality, only in so far as this thought can be realized in a possible experience as what is, at one stroke, the *ratio essendi* both of an experience in general and of its object. And when, in the course of critiquing pure reason, we see that our purely intellectual thought of an object in general is the intellectual form of our experience in general, we put thoughts of reflection to proper positive use in cognizing a priori the possibility of our a priori cognition. Because these thoughts have this proper positive use, our reason by its very nature cannot avoid finding the inferences of rational psychology appealing: the appeal these inferences enjoy is, in Kant's terminology, a transcendental illusion.

Let us return now to Kant's positive genetic characterization of a priori cognition, as cognition that 'our own cognitive capacity provides out of itself'.[126] As we saw, this is Kant's initial characterization of a priori cognition in the B-Edition Introduction. He provides it, moreover, in the course of raising the possibility that even 'our experiential cognition [*Erfahrungserkenntnis*]' is composed of 'that which we receive through impressions' and the cognitions that derive solely from our own cognitive capacity.[127] Why does Kant think that we must at least consider whether experiential

[126] B1. [127] Ibid.

cognition contains cognition that our capacity of cognition 'provides out of itself'? Kant's answer, it would seem, lies in his characterization of experience as cognition that our understanding works up from 'the raw material of sensible impressions' by comparing and connecting representations that objects have produced in our sensibility.[128] Connection, in Kant's sense, is a necessary relation. So the understanding must bring a cognition of necessary relation to its comparing and connecting of sensible representations, a cognition that is an essential constituent of any experience this comparing and connecting is to produce. On the assumption that the only other source of cognitions, other than experience, is our capacity of cognition itself, it follows that our experiential cognition contains cognition that our capacity of cognition provides out of itself.

A parallel discussion in the A-Edition Introduction confirms this interpretation. Immediately after providing his initial characterization of a priori cognition in terms of intrinsic clarity and certainty, Kant flatly claims that certain cognitions that 'must have an a priori origin' are contained within our experience, cognitions that 'perhaps serve only to establish connection among our representations of the senses'.[129] He then supports this claim with the observation that when we take away from experience 'all that belongs to sense' there remain 'certain original concepts and judgments developed out of them' that 'must arise completely a priori, independently of experience' because they 'seem to make it possible to say of objects that appear to the senses more than experience can teach us'.[130] In particular, these assertions 'contain true generality and strict necessity, the likes of which merely empirical cognition can never afford'.[131] The original concepts and judgments are, presumably, ones that our understanding employs in 'working on the raw material of sensible sensations' to produce experience, its 'first product'.[132]

Given how crucial this line of reasoning is to his critical philosophy, the fact that Kant rehearses it so quickly, and leaves it so sketchy, is initially rather puzzling. But that he does so is not surprising in light of the fact that it is one already familiar to his contemporaries from the work of Leibniz and his successors. Consider, in particular, that on Leibniz's account the representations through which we represent appearances as standing in

[128] B1. [129] A2. [130] Ibid. [131] Ibid.
[132] A1.

necessary relations are the thoughts of reflection, thoughts that derive solely from reflection, independently of the senses. We establish the real possibility of such necessary relations in general, in enjoying a purely a priori insight into the nature of our mind. But we also apply the thoughts of reflection to objects of our senses. Our perception presents us with appearances that we, through concepts such as *unity* and *substance*, understand as exhibiting particular intelligible essences. We can, for example, form distinct ideas of bodies in mechanistic terms and, through experiment, correlate our ideas of sensible qualities with these distinct ideas. To be sure, we cannot have insight into the appearances themselves, since being infinitely complex representations we cannot raise them to clarity and distinctness. But what justifies us in regarding the appearances as exhibiting real intelligible essences is the assurance that God, in his wisdom, creates the best possible world and that the best possible world is one in which created minds enjoy sensory perceptions that yield clear, if not distinct, ideas.[133]

When Kant suggests that these representations serve 'only to establish connection among our representations of the senses', he is raising the possibility that these representations yield cognition only of possible objects of experience. He is, thus, raising the possibility that, *pace* rational metaphysicians, these representations cannot be employed, in general metaphysics or ontology, to achieve insight into the possibility of things in general—let alone in special metaphysics to achieve insight into God, the soul, or the intelligible world. But if our ability to have such insight is denied, the question becomes pressing: how can we legitimately employ these genetically a priori representations to establish connections among the representations of our senses? If these representations do not provide us grounds for any genuine purely a priori insight, then it seems we have to turn to experience to establish the possibility of such connections. But experience *itself* does not yield any consciousness of necessary relations among appearances. So it seems that we lack the resources to establish the real possibility of the necessary relations we take appearances to stand in when we take them to provide us grounds of cognition. What is at stake is the very possibility of experience: for appearances constitute experience, empirical cognition, only if they provide us legitimate grounds for representing things in respect

[133] For a detailed reading of Leibniz's account of our knowledge of nature that develops this line of interpretation, see Donald Rutherford's *Leibniz and the Rational Order of Nature*, ch. 4 (esp. the section on the analysis of phenomena), and pt III.

of their specific identity, and they cannot provide us such grounds if we cannot establish the real possibility of their standing in these necessary relations.

Kant presents this problem about our ability to establish the real possibility of the necessary relations that we take appearances to stand in, in so far as they are to constitute experience, as a problem concerning the a priori objective validity of the categories in respect of appearances.[134] And, in the transition to the Transcendental Deduction of the Categories he presents his solution to this problem as follows:

> There are only two possible cases in which synthetic representation and its objects can come together, necessarily relate to each other, and, as it were, meet each other: Either if the object alone makes the representation possible, or if the representation alone makes the object possible. If it is the first, then this relation is only empirical, and the representation is never possible a priori. And this is the case with appearance in respect of that in it which belongs to sensation. But if it is the second, then since representation in itself (for we are not here talking about its causality by means of the will) does not produce its object as far as its existence is concerned, the representation is still determinant of the object a priori if it is possible through it alone to *cognize something as an object*.[135]

In other words, if they are to be possible our theoretical a priori cognitions of objects must comprise the *ratio essendi* of our cognition of objects: in no other way can they provide us genuine grounds for cognizing objects a priori, and thus grounds for determining necessary relations among objects. And, in particular, they cannot, as Leibniz held, provide us grounds for determining necessary relations among objects in virtue of comprising the *ratio essendi* of things in themselves. Kant advances this solution as the fruit of a genuine insight, a cognizing a priori of our theoretical a priori cognition from its *ratio essendi*. He is thereby claiming that at least some of our a priori cognitions are *first* a priori principles: principles, cognitions that are not only grounds of grounds that follow certain rules, but principles that, as a priori principles, ground themselves.

But what, then, is this genuine insight into the possibility of our a priori cognition to which Kant lays claim? In order to answer this question, we need to attend to Kant's characterization of the understanding, in the sense in which the understanding is to be contrasted with sensibility, as the mind's

[134] A89–90/B122–3. [135] A92/B124–5.

capacity to 'bring forth representations itself [*von selbt hervorzubringen*], or the spontaneity of cognition'.[136] Kant provides this characterization of the capacity of understanding at the opening of the Transcendental Logic in the service of specifying the task of this portion of the first critique—namely, that of explaining what genuine a priori cognition the capacity of understanding does and does not, of itself, ground a priori. As we will see, the genuine entirely a priori insight that Kant thinks we can have into the possibility of our theoretical a priori cognition is, in the first instance, insight into our higher cognitive capacity, or understanding. It is, in particular, insight into how this capacity, as a capacity of the spontaneity of our cognition, contains the categories, so that the genesis of these concepts in the exercise of this capacity, renders them '*self-thought* a priori first principles [*Principien*] of our cognition'.[137] Moreover, attending to this characterization of the understanding will shed light on Kant's more general genetic characterization of the a priori: the capacity of the spontaneity of cognition also grounds the possibility of our sensible theoretical a priori cognition, such as that of mathematics, in so far as such cognition consists in realizing an intellectual form in sensible matter in an impure exercise of the spontaneity of our cognition.

In characterizing the understanding as the spontaneity of cognition, Kant uses the term 'spontaneity' in a technical Leibnizian sense, one that needs to be understood in the context of certain Aristotelian conceptions of activity, capacity, and power. A power is that in virtue of which something acts—that is, constitutes a sufficient real ground (perhaps only given certain circumstances) of something's actually having some determination. Corresponding to every power is a capacity, or active potency: that in virtue of which a subject, in exercising a power, is active. According to Kant, a capacity is the inner possibility of a power. And activity is the inner action of a capacity, a striving (*Bestrebung, conatus*) through which that capacity actuates itself as a power. As we shall see, spontaneity, in Kant's account, is self-activity (*Selbsttätigkeit*), an activity with an inner principle sufficient to determine the power it realizes, a power that constitutes the being of the resulting subject of power.[138]

[136] A51/B75. [137] B167; Kant's emphasis.

[138] Kant offers these characterizations of activity (*Tätigkeit*), capacity (*Vermögen*), and power (*Kraft*) in his lectures on metaphysics: see, e.g., 28: 434; 28: 565; 27: 72. I discuss these characterizations in detail, and use them to clarify Kant's account of the spontaneity of cognition in 'Kant on Pure Apperception

Kant distinguishes two sorts of spontaneity, absolute and relative, or conditioned, spontaneity. Absolute spontaneity is a self-activity that itself suffices to determine its own inner principle, and so a self-activity in which an individual subject determines what power it realizes independently of that subject's interaction with any other subject, i.e., independently of any exercise of its receptivity (if it has a receptivity), and so solely in the exercise of its own capacity.[139] A conditioned spontaneity, in contrast, is one in which the inner principle of a subject's self-activity is not determined solely by that individual subject's exercise of its capacity. Kant offers as an example of a relative, or conditioned, self-activity, the activity in which a body, once it has been moved, realizes the motive power to continue in the same motion (unless impeded by another body).[140] This self-activity is conditioned, and not absolute, because its inner principle, and so what motive power that body exerts, is determined by the interaction with other bodies that set it in motion.[141] When Kant speaks of absolute spontaneity—as he does in speaking of the absolute spontaneity of cosmological, or transcendental, freedom—he typically has in mind an activity that determines what power it realizes independently of phenomenal causality, the causality that constitutes the being of things as they appear.[142] Indeed, when he speaks of spontaneity without qualification,

and the Spontaneity of Cognition' (unpublished MS). In invoking his characterizations of capacity and power to understand Kant's account of our cognitive capacities, I follow Beatrice Longuenesse. See her important *Kant and the Capacity to Judge* (Princeton, NJ: Princeton University Press, 1998), 7–8.

[139] 28: 267. [140] Ibid.

[141] See his (late 1770s) Politz lectures at 28.1: 267–8. Kant's characterization of inertia may strike the reader as inconsistent with his commitment to Newtonian physics. It is consistent, however, because Kant maintains that motive power—the power that a body takes on in actually being moved—is derivative, and not fundamental: it is to be explained by the way in which all bodies that constitute the physical world as such constitute a dynamical community in virtue of their standing in thoroughgoing relations of mutual interaction with each other through the fundamental, and so universal, attractive and repulsive moving powers of matter described in Newtonian physics. Kant holds that the notion of self-activity applies to bodies only in so far as we explain their motions in terms of motive powers.

[142] Note, in particular, that Kant's definition of transcendental freedom in the Antinomies specifies that it is a capacity that we have as subjects of understanding that belong to the sensible world and so think and choose in time: the transcendental concept of freedom, he says, is 'the capacity of beginning a state *from oneself* [*von selbst*], whose causality does not again stand under another cause according to the laws of nature' (A533/B561). Moreover, in the Antinomies, he also characterizes this freedom as one we exercise only as subjects of an absolute spontaneity. And in his lectures on metaphysics he specifies that the transcendental concept of freedom is that of absolute spontaneity in contrasting it with the concept of practical freedom (28.1: 267). What makes the activity in which we realize the capacity of transcendental freedom as a power an absolute self-activity, then, is its determining the power independently of any causality exercised by things as they appear. Note, too, how at R6077 in characterizing transcendental freedom as 'the absolute spontaneity to act', Kant contrasts this spontaneity

he typically has in mind absolute spontaneity, understood in this sense. Finally, it will prove important to see that an absolute self-activity need not be a complete, and so pure, spontaneity. In other words, a subject's absolute self-activity may, in realizing a power, essentially incorporate an exercise of its receptivity, and so its being affected (even its being affected in some particular way). Provided that a subject's exercise of its capacity to act, in determining what power it realizes, suffices of itself (independently of any exercise of that subject's receptivity) to determine what that power is in some respect, it is—if only to that extent—an absolute self-activity.[143]

In light of this sketch of Kant's conception of spontaneity, consider again Kant's characterization of the understanding as the capacity for the spontaneity of cognition.[144] The spontaneity in question is, I suggest,

with 'the *spontaneitas secundum aliquid*, since the subject is nonetheless *aliunde* determined through *causas physice influentes*' (18: 443).

[143] Here it is helpful to consider the case of an organism. According to Kant, our conception of an organism is that of the subject of an absolute, but incomplete and so impure, spontaneity. For it is the conception of a subject of an activity that realizes that organism's distinctive teleologically organized dynamical constitution in some matter (considered in its specific variety, the subject of some determinate moving power) only in incorporating the causality distinctive of that matter. An organism's activity in realizing the power that constitutes its being is self-activity only in so far as that organism's capacity to act of itself determines this power in respect of the distinctive teleologically organized dynamical constitution it realizes. That our conception of an organism is, on Kant's account, that of the subject of a spontaneity—and indeed, a self-activity that is absolute, relative to any phenomenal causality—is implicit in his contention, throughout the third critique, that organisms as such cannot be explained according to laws of nature. For laws of nature are laws that characterize a causality in which a subject's exercise of its capacity to act determines what power it realizes only in conjunction with its receptivity—in Kant's terminology, a mechanistic causality. Consider, in particular, that matter, considered simply as such, is the subject of the attractive moving power characterized by Newton's law of gravitation. And this law, as a law of mutual interaction, characterizes a power that is determined only by the conjunction of matter's capacity to act and its receptivity. And, of course, Kant holds that phenomenal causality is determined solely under laws of nature. It follows that, in Kant's account, an organism as such cannot be explained by laws of nature, in so far as it is the subject of a spontaneity that determines what power it realizes independently of the causality exercised by the matter that constitutes its body. Kant draws this conclusion himself when he claims that to conceive of a body as an organism is to conceive of it as the subject of spontaneity and thus to posit a ground of its distinctive dynamical constitution in the super-sensible ground of our experience (5: 411). Note that this passage is one in which Kant uses 'spontaneity' to refer to spontaneity that is—relative to phenomenal causality—absolute.

The reading of Kant's notions of spontaneity and absolute spontaneity that I am proposing is, to my knowledge, one that has yet to be considered by his commentators. It is also apt to be controversial. For example, it rejects Allison's equating of absolute spontaneity with transcendental freedom (See *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 15 and 60. As we will see, however, the two need to be distinguished: transcendental freedom is the absolute spontaneity of our thought and choice that makes theoretical and practical cognition possible for us. But I cannot defend my reading any further here. I argue for this reading in detail in 'Kant on Pure Apperception and the Spontaneity of Cognition' (unpublished MS).

[144] A51/B75.

absolute (relative to phenomenal causality): the spontaneity is one in which the capacity of understanding 'brings forth representations itself' by determining its own inner principle, independently of any operation of sensibility.[145] Thus, in Kant's account, the capacity of understanding is not—as has commonly been supposed by commentators—a capacity for just any activity of representing. It is one for an activity of representing in which the capacity for that activity of itself, and so independently of any exercise of its sensibility (its receptivity for representations), suffices to determine what power to cognize it realizes (at least in some respect). Moreover, this self-activity determines its own inner principle, a principle that characterizes it essentially, in representing this inner principle.

Kant's characterization of the understanding at A51/B75, then, is a highly rich and, in the eyes of many, highly contentious one. Indeed, one might object to my reading of this characterization on the grounds that it makes for a characterization too rich and contentious for Kant to have offered, as he does, without any elaboration or motivation. But this objection overlooks how this characterization of the understanding would, in fact, have been uncontroversial among Leibnizians, because it is one straightforwardly entailed by their conception of the understanding. Recall that the subject of thinking, on this conception, is not only a substance, but one whose essential activity makes manifest to that subject its nature as a thinking simple substance. Now in the Leibnizian tradition a substance, or *ens per se*, is as such the subject of an absolute spontaneity: every substance, as an individual thing, has a distinctive character, one that distinguishes the power that constitutes its substantial being; what makes a substance an *ens per se*, and an ultimate subject of predication, is its being the subject of a

[145] This reading clarifies Kant's remark, at A67/B92, that earlier he had explained the understanding 'only negatively, as a non-sensible capacity of cognition'. Kant is here referring back to his characterization of the understanding as the capacity for the spontaneity of cognition at A51/B75. (Here I disagree with Erdmann, who maintains that Kant was mistaken, for he had not previously given any merely negative characterization.) What makes this characterization merely negative is that it specifies the inner principles of this spontaneity merely as representations that the spontaneity determines independently of sensibility. It does not specify determinately the inner principles that this spontaneity determines of itself. Indeed, it is just this absence of positive determinate content that, according to Kant, belies traditional general metaphysics' claim to be practicing a genuine science of being in general (A238/B297–A247/B303). It is only when Kant has specified the particular representations that this capacity 'brings forth itself', and exhibited a priori how these representations are, as a priori principles in the from-grounds sense of 'a priori', what suffice collectively to characterize completely the inner possibility of our understanding, and thus constitute the nature of our capacity of understanding, that he has provided a positive explanation of the spontaneity of our cognition.

spontaneity that determines the distinctive character of the power it realizes of itself, and so independently of its interaction with any other being. What is distinctive of a substance that is a subject of understanding, a mind, is that its fundamental activity is a self-activity that gives its own common nature as a mind to itself immediately in an a priori cognition.

As I mentioned earlier, where Kant parts with the Leibnizian tradition is in denying that the self-activity that constitutes the mind as a subject of the power to cognize is one in which it enjoys an a priori cognition of itself *as a thing*. The unity of representations that a finite mind's self-activity realizes, through the categories, in realizing its power to cognize through its particular sensible representations is, of itself, not a property of a thing at all, let alone an essential property of a substance. Indeed, according to Kant, the first and most fundamental exercise of the spontaneity of our cognition is that in what he terms 'the transcendental synthesis of imagination', one that unites the manifold of given intuitions in accordance with the categories to realize in them what he terms 'the original synthetic unity of apperception'.[146] This unity of apperception, however, consists in the numerical identity of the spontaneity of one's cognition throughout any manifold of sensible intuition that this spontaneity could, in effecting the transcendental synthesis of imaginations, unite to produce cognition. And Kant holds that we can, through pure apperception, become conscious in general of the numerical identity of one's activity in the transcendental synthesis of the imagination as what constitutes one's numerical identity, throughout the manifold of given intuitions, as the subject of the power to cognize through this manifold. For this reason, the rational psychologist's mistake is an entirely natural one. Since the unity of representations we represent in the categories is in fact constitutive of the identity of the subject of cognition, it is all too easy to mistake this representation of unity as a cognition of the subject in respect of its identity as a thing in itself.[147]

[146] B151–2.

[147] According to Kant, a full critique of traditional metaphysics would also offer an error theory, along similar lines, of the doctrine of transcendental predicates—that is, of the properties of unity, truth, and goodness as essential properties common to all beings in general and thus as properties that transcend the different Aristotelian categories of being. These concepts are, he contends, merely 'logical requisites and criteria' that our reason puts to proper use only in striving to achieve a 'unity of comprehension' in our cognition of objects. He stresses, in particular, that traditional metaphysicians have 'carelessly' mistaken these concepts for properties of things in themselves, in conceiving of them as 'belonging to the possibility of things itself' (B114). In the Dialectic, Kant contends that these concepts differ from the categories in that they lack any genuine constitutive employment in cognition

For our purposes, it is instructive to see how, in the A-Edition Transcendental Deduction of the Categories, Kant advances 'the principle of the original synthetic unity of apperception', the claim that we are 'conscious a priori' of 'the thorough-going identity of ourselves with regard to all representations that can ever belong to our cognition, as a necessary condition of the possibility of all representations . . .'.[148] The consciousness in question is a general consciousness had, by drawing on pure apperception, of the spontaneity of our cognition as an activity that must, in its exercise in the transcendental synthesis of imagination, be numerically identical throughout the manifold of given intuition it unites in accordance with the categories—on pain of those intuitions being ones that could not belong to our cognition.[149] This necessary unity that the manifolds of given intuitions must have, under the original synthetic unity of apperception, in so far as they are to belong to our cognition is, in Kant's account, what most generally constitutes the necessary unity appearances have in so far as they constitute experience. The consciousness of this identity of the subject is a priori, because the consciousness of one's activity, the spontaneity of cognition, as numerically the same throughout all one's possible representations is registered in that consciousness as a ground of possibility of this activity's actuating a subject with the power to cognize through those representations, and so a subject of cognition that is numerically identical throughout those representations. It is because the exercise of our capacity of understanding that realizes it as a power is, in this way, essentially apperceptive that Kant characterizes our capacity of understanding as the capacity for the synthetic unity of apperception.[150] Note how, on the present reading, when Kant says that we are conscious of the identity of this subject a priori, he thus uses 'a priori' in the from-grounds sense.[151]

I am now finally in a position to sketch, if only in outline, how Kant maintains Leibniz's genetic characterization of the a priori, but develops an account of the genesis that produces a priori cognition sharply at odds

of objects; they are ideas, problematic concepts that our reason properly puts only to a regulative employment.

[148] A116. [149] A106–8. [150] B134 n.

[151] Notice that Kant says we are conscious a priori of this identity, not that we cognize it a priori: this reflects his doctrine that we only designate this subject transcendentally, a designation that does not distinguish it from any other possible subject, and that thus does not constitute a cognition. I develop and defend this reading of the conception of the identity of the subject in play in the Transcendental Deduction of the Categories in 'Unity of Apperception and Identity of the Subject', (unpublished MS).

with Leibniz's. Kant maintains this characterization because he holds that what makes a content an a priori cognition, a consciousness of a *ratio essendi* through which the subject can cognize something a priori in the from-grounds sense of 'a priori', is its originating in an exercise of the spontaneity of its cognition. An inner principle determined in this absolute spontaneity is one that characterizes essentially, as one of its *rationes essendi*, the power to cognize that this absolute spontaneity realizes. Moreover, the inner principle of an exercise of the spontaneity of cognition is essentially given in that spontaneity to its subject as such a principle, and so essentially constitutes an a priori cognition of the power that the absolute spontaneity realizes in the from-grounds sense of 'a priori'. Now, for Leibniz, what is cognized as thing through the power that is determined in the absolute spontaneity of cognition, just is the thing that thinks itself, because the thing that thinks, in so far as it is a thinking thing, simply consists in this power. Thus, in Leibniz's view, all the objects of our intellectual ideas, objects that are, as its inner a priori principles, given in and characterize the fundamental activity of a mind ('Being, Unity, Substance, Change, Duration, Action'), characterize its being as the thing that thinks. Kant, however, maintains against Leibniz that what makes the inner principles of an exercise of the spontaneity of our cognition (such as the categories) genuine a priori cognition is not the possibility of employing them in pure apperception. What makes them genuine a priori cognitions, rather, is their being concepts in accordance with which the transcendental synthesis of the imagination must unite the manifold of given intuitions, if the spontaneity of our cognition in this synthesis is to actuate any power to cognize objects at all. It is thus only as inner principles of this impure exercise of the spontaneity of our cognition that the categories constitute genuine a priori cognition. So, in Kant's account, the categories do constitute genuine a priori cognitions—in the from-grounds sense of 'a priori'—of things as such, that is, of things in respect of their objective reality. But they constitute such cognitions only of things that are possible objects of the intuitions that are combined in the transcendental synthesis of the imagination. They do not, *pace* Leibniz, constitute genuine a priori cognitions of whatever thing, or things, that are the subject of the spontaneity of our cognition.

Kant maintains that certain sensible contents are fixed simply by the forms of our sensible intuition, and so grounded solely in the nature of our sensibility, in so far as this nature conditions the possibility of our

cognition. But he holds that these contents constitute purely sensible a priori cognitions, only as they condition (under the principle of the original synthetic unity of apperception) the possibility of the exercise of the spontaneity of our cognition that realizes us as subjects of the power to cognize objects, and so as they originate in the transcendental synthesis of the imagination. His idea is that this synthesis, as one of the imagination, must be 'figurative':[152] it must consist in an essentially apperceived act of generating, successively and so part by part, a figure in thought according to a concept that is a determination of the category of quantity.[153] In so far as it is figurative, then, the possibility of the transcendental synthesis of the imagination, and so the possibility of the spontaneity of our cognition's realizing a genuine power to cognize objects, numbers among its *ratio essendi* the inner principles of an exercise of the spontaneity of our cognition that determines these principles in virtue of incorporating the forms of our sensible intuition. In short, what makes our purely sensible, as well as our purely intellectual, a priori cognition a priori cognition is its originating in the transcendental synthesis of the imagination. And what, in general, lies behind Kant's genetic characterization of a priori cognition is the contention that all of our genuine a priori cognition either consists of, or derives from, certain fundamental a priori principles, all of which, in turn, are a priori principles of our cognition ultimately in virtue of characterizing, under the principle of the original synthetic unity of apperception, the inner possibility of our power to cognize objects.

This reading of Kant's genetic characterization of a priori cognition, in which it is a corollary of his conception of the understanding as the spontaneity of cognition, and of its relation to Leibniz's antecedent views, finds some confirmation in Kant's famous claim about the original acquisition of a priori cognition:

> The *Critique* admits absolutely no divinely implanted (*anerschaffene*) or innate (*angeborne*) *representations*. It regards them all, whether they belong to intuition or to concepts of the understanding, as *acquired*. There is, however, an original acquisition (as the teachers of natural right formulate it), consequently also of that which previously did not exist, and therefore did not pertain to anything before the act. Such is, as the *Critique* shows, first of all, the form of things in space and time, and secondly, the synthetic unity of the manifold in concepts; for neither of

[152] B150. [153] B154.

these is derived by our capacity of cognition from the objects given to it as they are in themselves, but rather it brings them out of itself a priori.[154]

This doctrine of original acquisition clearly develops the genetic conception of the a priori, and in a way that links this conception to the conception of the understanding as a capacity for the spontaneity of cognition. What is more, this doctrine states elegantly Kant's main complaint with the Leibnizian way of understanding these conceptions of the a priori and the understanding. When he maintains that the categories, and more generally all the a priori principles of our cognition, are essentially derived in an original acquisition, Kant is maintaining that it is not as a representation of an a priori principle that characterizes the nature of things in themselves, and so in virtue of registering in cognition *rationes essendi* as they determine the real possibility of a thing in itself (in the Leibnizian view, divinely implanted dispositions that are properties of a thing in itself whose nature is to think), that they realize our power to cognize objects and constitute a priori principles. He is, in other words, insisting *pace* Leibniz and others that all the representations that originate in exercises of the spontaneity of our cognition—including those grounded solely in, and expressive of the nature of, our sensibility—constitute a priori principles of our cognition merely as representations.[155]

The project of the first critique is to provide a touchstone for distinguishing our genuine a priori theoretical cognition, if we have any, from concepts that only seem to us to constitute such cognition. Moreover, the positive part of this project requires establishing the real possibility of human a priori theoretical cognition. This, in turn, requires most fundamentally cognizing a priori the real possibility of the first a priori principles of our cognition, in respect of how they ground their own real possibility and, thereby, the possibility of all other a priori cognition. Indeed, Kant hopes to do this in a way that demarcates, in respect of their different kinds, all

[154] *On a Discovery*, 8: 221.

[155] This reading also sheds some light on Kant's characterization of the categories as '*self-thought* a priori first principles [*Principien*] of our cognition' (B167; Kant's emphasis). He offers this as an alternative to the Leibnizian view, which he disparages as a *deus ex machina*. He stresses 'self-thought' because it is in being self-thought and so realized as thoughts (and not merely as ideas, dispositions of the soul that, in a mind, can be registered as such in self-cognition) that the categories constitute the first a priori principles of our cognition. In other words, his point is that the categories constitute the first a priori principles of our cognition only as representations, not—as on the Leibnizian view—as properties of a thing.

the a priori principles of our cognition—which amounts to providing an exhaustive taxonomy of all the genuine sciences of human cognition.

The Transcendental Analytic purports to execute this positive project in two stages. The first stage, executed in the Analytic of Concepts, exhibits a priori, in the from-grounds sense of 'a priori', the real possibility of the categories' constituting genuine first a priori principles of our cognition, as inner principles that the exercise of the spontaneity of our cognition in the transcendental synthesis of the imagination determines solely out of the nature of our capacity for the spontaneity of our cognition. And this amounts to exhibiting how, under the principle of the original synthetic unity of apperception, all of these a priori principles of our cognition constitute the most fundamental *rationes essendi* that determine, in virtue of their origin in the transcendental synthesis of imagination, the real possibility of experience in general in respect of its purely intellectual form. In the Analytic of Principles, Kant executes the second stage—namely, exhibiting a priori all the more determinate a priori principles that, given the form of our inner sense, determine the possibility of our experience in general. These transcendental principles (the transcendental schemata and the principles of pure understanding) are what collectively make possible the application of the various categories in a possible experience in general. It is only in the second stage, according to Kant, that we cognize a priori how all the intellectual principles grounded in the nature of our capacity for the spontaneity of cognition itself, suffice collectively to determine the possibility of experience in general, in respect of its intellectual form. Notice how, on the present reading, Kant's genetic characterization of the a priori serves as the real definition in terms of which the Transcendental Analytic as a whole exhibits a priori the real possibility of our a priori cognition.[156]

The readings I have proposed in this section of Kant's characterization of the spontaneity of cognition, and of his corollary genetic characterization of a priori cognition, are, as far I know, original and likely to be controversial. A full development and defense of these readings would require showing, in detail, how they illumine the account of human cognition, including in particular human theoretical a priori cognition, that he develops in

[156] This exhibiting is what Kant describes in the Discipline of Pure Reason as 'the use of reason through construction of concepts' (A723/B751–2.), and what is so exhibited is, in his account, the 'rational cognition from concepts which is called philosophical' (A724/B752).

the first critique. We have already begun to see, if only in outline, how these readings shed light on his account of the original synthetic unity of apperception, his theory of synthesis, and his account of the nature of his project of critiquing pure reason. Elsewhere,[157] I have explained in detail how my reading of his conception of the spontaneity of cognition sheds light on his doctrines concerning the original synthetic unity of apperception. And in other work I hope to explain in more detail how my readings of his conceptions of apriority and spontaneity help clarify both Kant's account of his project of critiquing pure reason and his theory of synthesis, the exercise of the spontaneity of cognition that unites sensible intuitions to produce cognition. Let me close the present chapter with a few brief remarks about these last two sources of evidence for my readings of apriority and spontaneity.

In the case of Kant's theory of synthesis, the present readings of his characterization of the understanding as the capacity for the spontaneity of cognition and of his genetic characterization of apriority serve to bring to light Kant's grounds for his famous, and apparently radically undermotivated, claim that the unity sensibly given representations have in a cognition cannot be presented to our consciousness simply in the affection of our sensibility, but that the understanding must produce this unity itself, through its activity in synthesis. Appreciating Kant's conceptions of a priority and spontaneity puts us in a position to see what fundamentally motivates this claim—namely, Kant's view that, unless synthesis, as an effect of a function of the absolute spontaneity of cognition, produced this unity, we could not have the insight into the possibility of experience that we must be able to have, if experience is to supply us a posteriori cognition.

In the case of the project of critiquing pure reason, these readings put us in the position to see how and why Kant thought that his project of providing an a priori proof of the possibility of our a priori cognition has as its subject-matter our capacity of pure reason—that is, the capacity of reason 'that contains the principles for cognizing absolutely a priori'.[158] They also help explain his otherwise puzzlingly sanguine claims about how what is contained in our reason cannot be hidden from view, and his seemingly excessive demand for architectonic structure: in his view,

[157] 'Unity of Apperception and the Spontaneity of Cognition', (unpublished MS). [158] B24.

the only assurance of the real possibility of our a priori cognition that critique can possibly provide requires essential definitions in Leibniz's sense—definitions that characterize how certain purely intellectual a priori cognitions serve collectively to determine the inner possibility of pure reason, and so of our all a priori cognition in general.[159]

[159] I am grateful to the following people for valuable discussions and/or comments: Robert M. Adams, Karl Ameriks, Tyler Burge, John Carriero, Tom Christiano, Suzanne Dovi, Barbara Herman, Lee Hardy, Des Hogan, Larry Jorgensen, Keith Lehrer, and Joseph Tolliver. Portions of this material were presented at a Kant workshop held at Notre Dame in the Summer of 2005, at the Brazil International Kant Congress, and to the graduate seminar I held at the University of Arizona in the fall of 2007: thanks to these audiences for helpful discussions.

6

Moral Necessity in Leibniz's Account of Human Freedom

R. C. SLEIGH, JR.

The topic of this paper is Leibniz's account of the nature of human free choice, and, specifically, the place of the notion of moral necessity in that account. We are fortunate to have available the work of Wolfgang Hubener, Sven Knebel, and Michael Murray, who have traced important aspects of the notion of moral necessity, as it occurs in Leibniz's study of human freedom, to their source in the thinking of certain seventeenth-century Spanish Jesuit philosophers. Hubener and Knebel have illuminated the work of the relevant Spanish Jesuits in a number of studies, and applied their findings to the elucidation of Leibniz's thinking on freedom and necessity. Michael Murray, in a series of important papers, has carried the project forward with considerable success.[1]

There are a number of advantages to their approach in virtue of its intimate connection with problems in philosophical theology that were as important to Leibniz as to the Spanish Jesuits. Perhaps the lead advantage is this: in the relevant tradition, which includes the work of St Thomas, Scotus, Molina, Bañez, and Suárez, there are well worn paths connecting theses about divine foreknowledge, the problem of the author of sin, the necessity of human free choice for moral responsibility, and the necessity of the latter, with various competing accounts of human free choice. For example, it seems to me unlikely that Molina would have reached the conclusions he did about the necessity of appealing to *scientia media*

[1] See, for example, Sven Knebel, 'Necessitas Moralis ad Optimum', *Studia Leibnitiana*, 23 (1991), 3–24; W. Hubener, 'Notio completa: Die theologischen Voraussetzungen von Leibniz' Postulate der Unbeweisbarkeit der Existentialsatz und die Idee des logischen Formalism', *Studia Leibnitiana*, Sonderheft, 15 (1988); Michael Murray, 'Pre-Leibnizian Moral Necessity', *Leibniz Review* (2004), 1–28.

in order to account for God's omniscience, but for his belief that only a libertarian account of human freedom will suffice, coupled with his belief that even divine causation of a person's choice is inconsistent with the libertarian position required. There can be little doubt that Leibniz's considered, fixed view about exactly what conditions must be satisfied for a person's choice to be free is not transparently obvious from the texts. So it is a considerable advantage to have worked out for us various connections among the relevant theses in philosophical theology and the notion of human freedom. There is the prospect that an examination of Leibniz's thinking with respect to the relevant aspects of philosophical theology will shed new and surprising light on his notion of human freedom. And, indeed, Hubener, Knebel, and Murray—especially the latter—go further and claim that Leibniz's considered opinion on these matters involved incorporating a particular line developed within the Spanish Jesuit tradition into his own account of free choice.

I begin with a short, compact discussion of the way in which various positions in philosophical theology relate to views about the proper account of human freedom. Then I outline the sophisticated, nuanced account of Leibniz's thinking concerning free choice, final causation, moral necessity, and related matters in philosophical theology, including his ideas concerning divine foreknowledge and the problem of the author of sin, developed recently by Hubener, Knebel, and Murray. I then outline an unsophisticated, unnuanced, traditional account of the same material, which I have on occasion defended. I really do not know whether either of these accounts is correct, or, if not, what alternative account is correct, although I wish I did. Nevertheless I argue that, all things in my ken considered, there seems to me to be insufficient reason currently available for abandoning the relative simplicity and familiarity of the unsophisticated, unnuanced position to which I have become attached.

I

In this section I take note of the way in which God's knowledge of free choices and God's relation to sinful choices were understood among those whose influence on Leibniz's account cannot be disputed.

In his writings, which purport to elucidate, but never contradict or even seriously amend St Thomas's teaching, the Dominican Bañez insisted that while St Thomas sometimes claimed that God knows future contingents, including free choices of creatures, in virtue of their being present to Him in His eternity, on Thomas's deepest account of God's knowledge of free choices of creatures, such knowledge is ultimately derived from His knowledge of His own will, His own decrees concerning creation. Bañez, and his Jesuit opponents, emphasized God's independence from creatures, applying this emphasis to God's knowledge by insisting that God's knowledge is ultimately self-knowledge.

Bañez's view has a wonderful simplicity to it: God's knowledge of necessary truths is ascribed to his comprehension of His own intellect, on which necessary truths are said to depend; and God's knowledge of contingent truths is based on His knowledge of His own will. According to Bañez, God's knowledge of actual free choices of creatures is derived from His knowledge of His actual will, while His knowledge of counterfactuals of freedom—counterfactual conditionals asserting what a creature would choose in various counterfactual conditions—is based on His knowledge of what He would have willed, had those circumstances been actual.

Unfortunately, as is so often the case in philosophy, simplicity comes at a price. The price comes about in the following way. An obvious non-occasionalist account of how God might know what a creature will freely choose on a given occasion would be this: what a creature will freely choose in the situation at hand is causally determined by natural circumstances then prevailing, both external and internal to the agent. But both the circumstances and the relevant causal laws are products of God's productive will. Of course, an occasionalist account would be even simpler. But those who concern us—for example Thomas, Scotus, Molina, Bañez, and Suárez—all rejected occasionalism. And they also rejected a crucial step in the explanation offered above—namely, the idea that a choice might be both free and yet causally determined by natural circumstances then prevailing. Each contrasted the way in which non-rational agents, sentient animals, and inanimate objects, for example, were determined to one outcome—determined *ad unum*—as it was put, by the natural circumstances prevailing, with the case of rational agents, which, when free, are not so determined to one outcome. Of course, where there is a choice, there is an outcome. But the idea is that in the natural circumstances

then prevailing another outcome was possible, i.e., without the need for divine intervention in the miraculous mode, and without abrogating any natural law, there yet might have been a different choice made—that is, elicited by the agent, as it was put.

Hence, Dominicans like Bañez were required to take a different tack in order to maintain the thesis that God's knowledge of the free choices of creatures is based on a knowledge of His own will, His own decrees concerning creation. The route taken was based on a particular understanding of the doctrine of divine concurrence. All of those who concern us presently accepted a strong form of the doctrine of divine concurrence, according to which every action of every creature is a joint effort involving input from both God and creature, with the result that numerically one and the same action is both an act of a creature and an act of God, although God and creature act in virtue of separate powers, and, hence, may be responsible for different features of the action. In the case of a free choice of a creature, Dominicans, like Bañez, claimed that God's contribution, without which there would be no such choice, includes His moving the creature to choice via a divine premotion of the creature's will, a divine physical predetermination, as it was sometimes put. And Bañez claimed that God's knowledge of this premotion, which was an exercise of God's will, sufficed to account for God's certain, infallible, a priori knowledge of the actual free choices of creatures. And God's knowledge of counterfactuals of freedom, in this account, consists in His knowledge of what premotion He would have offered had the conditions noted in the antecedent of the relevant conditional been actual.

There is a problem concerning whether the account offered by Bañez provides for a satisfactory resolution to the problem of the author of sin. So we need to say something about the problem of the author of sin. It is a commonplace in treatments of the problem of evil to make as much use of the greater good defense as circumstances will permit. The basic idea of the greater good defense is this: in the case of at least some varieties of evils it is morally permissible for an agent A to permit, and even cause, an evil e to obtain, if A does so in order to bring about some good g such that A cannot bring about g's obtaining without e's obtaining, and it is better that both g and e obtain than that neither does. Commonplace as maximum use of a defense based on this claim was among those under consideration, equally common was the claim that permission or causation of sin could not be

justified in this way, either with respect to us or with respect to God. And it was a commonplace among Bañez's critics to claim that his account of divine foreknowledge of free choices that were sinful had the consequence that God was causally involved in the sins of creatures in ways inconsistent with His holiness. The response of supporters of Bañez was to attenuate the connection between God's premotion of the will of a creature in a sinful free choice and the sinful content of that choice. The resulting tension in Bañez's theory is obvious: the weaker the connection between God's premotion and the content of the resulting choice, the better for solving the problem of the author of sin, but the worse for accounting for God's certain, infallible, a priori knowledge of that content.

The Jesuits, Molina, and Suárez, rejected Bañez's premotion theory as an account of concurrence and also because of its implications concerning free choice. They saw premotion theory as a theory according to which God's role in each human action involves both moving the agent to action and, in addition, acting with the creature. The latter, according to Molina and Suárez, exhausted God's role as far as general concurrence is concerned. And they held that a premotion adequate to account for God's certain, infallible, a priori knowledge of free choices must determine the agent *ad unum*. Neither Molina nor Suárez was prepared to accept the idea that divine causal determination of a creature's choice is consistent with freedom, while natural causal determination is not.

Scotus, Molina, and Suárez offered pure libertarian theories, holding that if a choice is free, then, with all the circumstances then obtaining, natural and supernatural (at least with respect to God's ordinary concurrence), the agent can elicit a choice or not, elicit a choice specified thus, or otherwise. And they traced what they took to be the error of Thomas and Bañez to their acceptance of the principle that in every case in which an entity is reduced from potentiality in a certain respect to actuality in that respect, it is so reduced by something other than itself. The heart of the libertarian theory that they offered is the claim that free choices constitute counter-examples to this principle. And, indeed, it is fair to say that they took free choices to constitute counter-examples to the principle of sufficient reason in a form in which Leibniz took it to be a basic truth of metaphysics.

The libertarian theory offered by Molina and Suárez, taken in conjunction with theories of divine concurrence they proposed, fit perfectly

with the basic incompatibilist tendencies they shared—at least with respect to natural circumstances—with Thomas and Bañez; moreover, they are tailor-made to disarm the problem of the author of sin. But, of course, all these benefits come at a price. Leave aside for now the metaphysical price to which Leibniz, for one, would draw our attention—the abandonment of the principle of sufficient reason. The price on the theological side of the ledger also appears steep. The libertarian theory offered appears to loosen the bonds that tie creature to God in ways many saw as inconsistent with an acceptable account of divine providence. And then there is the problem of accounting for God's certain, infallible, a priori knowledge of free choices. 'Scientia media', the name of the key ingredient in the theory offered by Molina, was taken by critics to be the name of a problem, not its solution.[2]

Not surprisingly, various efforts to forge a compromise, or, at any rate, a more adequate account are to be found in the work of seventeenth-century theologians. Since our focus is on the notion of moral necessity in Leibniz's account of human freedom, I turn to a brief consideration of the views of those who might be called 'moral necessitarians', many of whom were Spanish Jesuits, e.g., Ruiz de Montoya, Diego Granado, and Sebastian lzquierdo. For a detailed and insightful study of the moral necessitarians, I recommend Michael Murray's 'Pre-Leibnizian Moral Necessity', *Leibniz Review* (2004), 1–28.

Consider the following passage presented by Murray from Izquierdo, which aims to locate the notion of moral necessity:

a subject has a metaphysical necessity to act when . . . if it failed to happen, two contradictories would be given, which is certainly repugnant. Something is physically necessary, however, when it could not fail to happen naturally and without a miracle, even if it could happen miraculously. Thus, finally, something is morally necessary when, by way of inclination, that which usually, or always,

[2] For Molina's account of matters, see Luis de Molina, *Liberi arbitrii cum gratia donis, divina praescientia, providentia, praedestinatione et reprobatione concordia*, ed. Johann Rabeneck (Oña and Madrid, 1953); Part 4 of the Concordia is translated into English by Alfred J. Freddoso: de Molina, *On Divine Foreknowledge* (lthaca, NY and London: Comell University Press, 1988). For Bañez's contribution, see Dominic Bañez, Tractatus de vera et legitima concordia liberi arbitrii creati cum auxiliis gratiae Dei efficaciter moventis himanam voluntatem, in *Commentarios inedites a la prima secundae de santo Tomas*, 3, ed. V. B. de Heredia (Madrid, 1948), 351–420. For Suárez, see Disputation 19 of Francisco Suárez, *Disputationes Metaphysicae*, repr. (Hildesheim: Georg Olms, 1965); Disputation 19 is translated into English again by Freddoso: Francisco Suárez, *On Efficient Causality* (New Haven, Conn. and London: Yale University Press, 1994).

or almost always, is accustomed to occur, cannot fail to happen, even if it can fail absolutely or in the light of a law of nature.[3]

There are a variety of views concerning human freedom that make use of the notion of moral necessity that Izquierdo set out to explain in the passage cited. Here is a quick summary of one that will serve our purposes. Whenever a human agent A makes a choice C, circumstances R obtain, consisting of motives, desires, beliefs, etc., of the agent, that morally necessitate the choice made, but, if the choice is free, do not metaphysically or causally necessitate that choice. To say that R morally necessitates C involves at least these claims: (a) the obtaining of R is consistent with A retaining the power to choose otherwise than C; and (b) the state of affairs consisting of R obtaining while C does not is neither contradictory (so R does not metaphysically necessitate C) nor miraculous (so R does not causally necessitate C); and (c) none the less, there is a relation of counterfactual implication between the obtaining of R and the agent's eliciting C such that, were God to know that R obtains, God would then have certain, infallible, a priori knowledge that C obtains.

That, in a nutshell, is the theory. I leave a consideration of its merits to others. I turn to a consideration of the claim that it—or something very much like it—is the account of human free choice Leibniz came to accept.

II

I am sympathetic to the study of Leibniz's philosophical theology on a number of accounts. First, it is an interesting topic in its own right; philosophical theology is a subject on which Leibniz focused his considerable talents with some frequency. Second, Leibniz's lifelong interest in various church reunion projects provides a complicating dimension to an understanding of his work on issues in philosophical theology, requiring considerable delicacy in interpreting what he wrote. A case in point is Leibniz's *Systema Theologicum*, a work ostensibly defending various theological positions associated with Roman Catholicism. And, thirdly, as mentioned previously, there is the reasonable expectation that we may employ a

[3] Trans. Michael Murray, ibid., 14.

knowledge of the then current problematic in philosophical theology, including Leibniz's contributions thereto, in order to ascertain his views on philosophical matters such as free choice. Nonetheless, I think that it is crucial to bear in mind the extent to which Leibniz inclined to put metaphysics first, and, taking care that nothing in his metaphysics contradicted revealed Christian theology, kneading and shaping theological claims so they fit his metaphysics. So in this section I remind you of some elements of Leibniz's metaphysics, to which he was committed, early and late, which seem to me to undercut, or, at any rate, significantly attenuate any serious motivation for him to adopt the views of the moral necessitarians.

There are two principles in particular that seem to me to bear in a special way on the questions that confront us. Each is a principle that Leibniz accepted early and late. The first is what I have called the 'doctrine of superintrinsicalness'; the second, the principle of spontaneity. In Leibniz's usage, an individual substance s has a property f intrinsically, just in case s has f, and were s to lack f then s would not have existed in the first place. On occasion Leibniz formulated the defining condition this way: s has f, and, for any x, were x to lack f then x would not be s. The doctrine of superintrinsicalness is the striking claim that, for any individual substance s and property f, if s has f, then s has f intrinsically.[4] In the neighborhood is the 'doctrine of superessentialism', which we may characterize as follows. Let us say that individual substance s has property f essentially just in case, s has f, and it is metaphysically necessary that if s exists, then s has f. Superessentialism is the doctrine that, for any individual substance s and property f, if s has f then s has f essentially. There are disputed questions lurking here. One is whether superintrinsicalness entails superessentialism. Another is whether Leibniz thought so, and, hence, accepted superessentialism. I wish to avoid these questions.

Leibniz, in typical seventeenth-century fashion, characterized the individual concept of an individual substance as that concept that contains all and only the properties that substance has, or would have, were it to exist, intrinsically. Thus, in virtue of accepting superintrinsicalness, Leibniz reached the atypical conclusion that each concept of each individual substance is complete, i.e., contains all and only the properties that substance

[4] See, for example, Gottfried Wilhelm Leibniz, *Samtliche Schriften und Briefe*, Deutsche Akademie der Wissenschaften (Darmstadt: Akademie-Verlag, 1923–), ser. 6, iv. 1645. Subsequent references to this edition will be by series and volume, e.g., A/6/4/1645. See also *Discourse on Metaphysics*, #30.

has, or would have, were it to exist. And Leibniz held views that have the consequence that the set of all individual concepts may be partitioned into an infinite collection of sets of compossible individual concepts. Thus, no such concept is in more than one collection, and each such set of concepts represents a possible world. Leibniz sometimes pictured divine creation as consisting in a single, world-actualizing, divine decree: for short—let this world be, or, more accurately, if less poetically, let this maximal set of compossible concepts be instantiated.

Consider this passage from Leibniz's letters to Arnauld: 'Every present state of a substance occurs to it spontaneously, and is only a consequence of its preceding state.'[5] Remarks in the textual neighborhood strongly suggest that the kind of consequence Leibniz had in mind is a causal consequence. Later, again in the correspondence with Arnauld, Leibniz offered a slightly more carefully crafted version of the Principle of Spontaneity:

> Everything happens in each substance in consequence of the first state that God gave it in creating it, and extraordinary concourse aside, His ordinary concourse consists simply in the conservation of the substance itself, in conformity with its preceding state and with the changes it [the preceding state] carries with it.[6]

So the Principle of Spontaneity, which Leibniz accepted, amounts to this:

For any individual substance s and state f of s, either f is a (causal) consequence of a preceding state of s, or, f obtains miraculously.

Leibniz supplemented the Principle of Spontaneity with the claim that all real causation, divine causality aside, is intra-substantial. Like many seventeenth-century philosophers—Malebranche comes to mind—he recognized forms of what we might call quasi-causality, which are distinct from real causality, and which hold both within a substance and among substances.

So much for the relevant metaphysical background. Given that background, at least as I have outlined it, I should think that, prima facie, the claim that Leibniz was a causal determinist—up to the point of miracles—and compatibilist, would look promising. Consider some free choice c of some agent a. According to superintrinsicalness, a would never even have existed, had a not elicited choice c. If one is willing to accept that level of determination as being compatible with freedom, as Leibniz did,

[5] Leibniz–Arnauld correspondence, as presented by Gerhardt in *Die philosophischen von G. W. Leibniz*, ii. 47.

[6] Ibid., ii. 91–2.

causal determination, either real or quasi, looks benign by comparison. And with the obvious caveat for divine intervention in the miraculous mode, acceptance of the principle of spontaneity strongly suggests that Leibniz was a causal determinist, committed to the thesis that each choice of an agent, including those that are free, is causally determined by the preceding state of that agent. But, of course, all of this is just 'prima facie', and up for interpretation. Those committed to the idea that Leibniz was a 'moral necessitarian' will balk at the suggestion that it is correct to construe the notion of consequence, as it occurs in Leibniz's account of spontaneity, as referring exclusively to a notion of a causal consequence. They will want to include consequences supported by appropriate counterfactual implications.

In the next section I provide one example of the kind of consideration that keeps me—perhaps blindly—from buying into the account Murray provides.

III

Consider the crucial case of counterfactuals of freedom. It was seen as crucial, in part, because Thomas's idea that God's knowledge of actual future contingents depended upon the presence of the relevant entities to God in His eternity was viewed with considerable reservation by all parties to these disputes, and, it had no straightforward application to counterfactuals of freedom. Another reason is that knowledge of such conditionals seems to be attributed to God in the Bible, the passages most often mentioned being at 1 Samuel 23: 7–13. Lastly, and most significantly, God's knowledge of such conditionals was taken by the parties that concern us—Dominicans, Molinists, and moral necessitarians—as essential to divine decisions concerning creation. Consider two possible world initial segments which are exactly alike, leaving aside possible divine activity, up to the point at which some creature A elicits a free choice. One initial segment is a segment of a world in which A elicits C; the other, one in which A elicits C*, (C and C* distinct). How does God know which continuation will occur, were He to create the relevant initial segment? The Dominican answer is that God knows in virtue of knowing what premotion of A's will He would have offered. This account was summarized by claiming that God's knowledge of the relevant counterfactual of freedom

is post-volitional. Both the Molinists and moral necessitarians rejected this answer, claiming that God's knowledge must be pre-volitional, and so not based on a knowledge of what He would have willed in the relevant circumstances, because, were it so based, freedom would be precluded from the creature's choice.

Michael Murray, in a patient, careful analysis of Leibniz's important piece, 'De Libertate, Fato, Gratia Dei', has noted a shift in Leibniz's position, which he takes to be a shift from a post-volitional account of God's knowledge of counterfactuals of freedom to a pre-volitional account.[7] The pre-volitional account is common to both Molinists and moral necessitarians, but, as Murray points out, Leibniz explicitly and frequently divorced himself from the Molinist position on the grounds, previously noted, that it is incompatible with the principle of sufficient reason. Murray notes that the position of the moral necessitarians is inconsistent with the principle of sufficient reason only if that principle is construed as requiring not only a reason, but a cause, in every relevant case. And, Murray concludes, we have no sufficient reason for so construing Leibniz's construal of the principle of sufficient reason.

I want to explain why I am not fully persuaded by these considerations. Writing about counterfactuals in a reading note, Leibniz began as follows:

> Generally, future conditionals are senseless things. Presumably when I seek to know what would have happened, had Peter not denied Christ, I seek to know what would have happened, had Peter not been Peter, because denying is contained in the complete concept of Peter.[8]

I think that this is Leibniz's official position. He was surely committed to the thesis that God has certain, infallible, a priori knowledge with respect to every proposition that has a truth-value. It's just that, given superintrinsicalness, counterfactuals, including, in particular, counterfactuals of freedom, are not such, according to Leibniz's official position. I am aware that in a variety of texts from different periods of his life, Leibniz provided various accounts of God's knowledge of a variety of relevant

[7] See Murray, 'Spontaneity and Freedom in Leibniz', in Donald Rutherford and J. A. Cover (eds.), *Leibniz: Nature and Freedom* (Oxford: Oxford University Press, 2005), 194–216.

[8] Leibniz, *Textes inédits*, ed. Gaston Grua (2 vols.; Paris: Presses universitaires de France, 1948; repr. New York: Garland, 1985), 358. Referred to hereafter as Grua.

items, including counterfactuals, and, hence, presumably, counterfactuals of freedom.[9] And, indeed, Leibniz criticized those who purported to find a difficulty in the idea that God had such knowledge. For example, in 'Rationale fidei Catholicae', after noting that the principle of sufficient reason applies to choices as well as every thing else, Leibniz wrote:

> it is easily seen how God would know what some Mind would choose, were it to come into some state, which, in fact, never will be actual, for it is not the case that God cannot see on what side the reasons would be more plausible or the emotions stronger. Of those who do not wish to make use of so natural an explanation on account of certain prejudices of the school, some are forced to doubt whether God can know such a thing, which, nevertheless, is unworthy of God, and is contrary to sacred Scriptures.[10]

Some of the accounts that Leibniz offered utilize the notion of possible worlds in ways that sound quite contemporary. Each of them seems on its face to have a failure of fit with Leibniz's official account of the status of counterfactual conditionals. I believe that all these offerings are what I have called replacement analyses, where Leibniz's intention is not to capture the ordinary meaning of what is being analyzed, but rather either to explain what is actually going on in reality, contrary to what those deprived of the true metaphysics may think, when the analyzed item is ordinarily employed, or, at a minimum, to provide an alternative reading that is in the neighborhood of the ordinary meaning, but that is not flatly inconsistent with the true metaphysics. Indeed, the reading note quoted above continues as follows:

> Nevertheless, it is excusable that, on this occasion, by the name Peter is understood what is involved in those things [attributes of Peter] from which the denial does not follow, and at the same time there is subtracted from the entire universe all those things *from which it does follow*. And then sometimes it can happen that a decision follows per se from the remaining things posited in the universe. But sometimes it does not follow unless a new divine decree occurs based on the rule of the best. If there is no natural chain or succession from the remaining things posited, then it is not possible to know what will happen except on the basis of a decree of God in accord with what is best. Therefore, the matter depends either

[9] For a helpful summary of the various alternatives Leibniz considered, see Michael Griffin, 'Leibniz on God's Knowledge of Counterfactuals', *Philosophical Review* (1999), 317–43.

[10] A/6/4/2318–19.

on the series of causes, or on a decree of the divine will. They do not seem to gain anything at all by means of middle knowledge.[11]

This passage makes clear that, even though on Leibniz's official account counterfactuals of freedom are strictly speaking meaningless, still a consideration of the various replacement analyses that he offered and discussions related thereto, may shed light, as Murray has suggested, on Leibniz's account of human freedom. Thus, in this reading note, after offering a general recipe for constructing a replacement analysis for a counterfactual, Leibniz noted that the type of analysis offered indicates that there is no need to utilize Molina's *scientia media* in order to account for God's knowledge. And we know that Leibniz saw *scientia media* as closely connected with Molina's libertarian account of free choice. By my lights, what we find in this replacement analysis tends away from a pre-volitional account of the relevant divine knowledge, and towards the post-volitional account he offered in his paper, entitled 'Scientia Media', where Leibniz summarized the position he then advocated as follows:

God knows future absolute things because He knows what He has decreed, and future conditionals because He knows what He would have decreed. Moreover, He knows what He would have decreed, because He knows what, in this case, would be the best, for He would decree the best. Were it otherwise, it would follow that God could not know for certain what He Himself would do in this case.[12]

'Scientia Media' was written by Leibniz prior to 'De Libertate, Fato, Gratia Dei', in which Michael Murray has located a switch in Leibniz's thinking from a post-volitional account to a pre-volitional account of God's knowledge of the relevant material. And, as noted, Michael Murray has connected this change to Leibniz's use of the terminology of the moral necessitarians in striking ways. Here is a passage from 'De Libertate, Fato, Gratia Dei', that contains the pre-volitional account that Michael Murray has highlighted for us:

God does not decide that Peter will sin; rather what God decides is that, on account of hidden principles of His wisdom, from infinitely many possible creatures, indeed,

[11] Griffin, 'Leibniz on God's Knowledge of Counterfactuals', 358. The English words 'from which it does follow' replace the Latin '*ex quibus non sequitur*'. I assume Leibniz lost track of things here.

[12] A/6/4/1374.

Peter is chosen, in whose concept, i.e., the absolutely perfect cognition that God has of him before He decided on his existence, there is contained the fact that, were he to exist, he would freely sin; that is, from infinitely many ways of creating the world, or, from infinitely many possible decrees of His, God has chosen those with which that series of possible things is connected, in which Peter's freely sinning is contained. Hence, in fact, God merely grants existence by an actual decree to possible Peter, who will sin. And so He does not decide that Peter sins, but merely that possible Peter is admitted to existence, even though he is going to sin. Someone who was not going to sin—granted freely—would not even be this Peter.[13]

Leibniz generalized the strategy involved in this passage in readings notes probably written after 'De Libertate, Fato, Gratia Dei'. Discussing God's role in the dispensation of grace, he wrote:

It must not be held that God decides specifically concerning aids [to salvation] given to Peter or to Judas; rather, He decides whether He wishes to admit to existence the possibles—this Peter, this Judas—with the total series of aids and circumstances already included in the complete concept of each. In fact, He does not even decide that, considered in itself, but rather whether He wishes to admit to existence the universal series of possibles in which Peter and Judas, endowed with the qualities stated, are contained among infinitely many others . . .

The object of the divine decree is not the man, but rather the total series of possibles making up this universe, taken with all its states, past, present, and future.[14]

The account offered in these writings is a permanent feature of Leibniz's thinking. His way of conceptualizing God's choices and decisions with respect to creation is on display here. There is an effort, based ultimately on superintrinsicalness, to place as much theodicean pressure on a single, world actualizing divine decree as possible. A consequence of this effort is this: in Leibniz's scheme, God's post-volitional knowledge of actuals is restricted to knowledge of what set of complete individual concepts is instantiated; knowledge of the features of the individuals thereby made actual is pre-volitional, just as Michael Murray affirms. But Leibniz's scheme is different from that presupposed by the scholastics under consideration in ways that obstruct inferences to theses concerning the conditions required for human freedom from premises about the nature of God's knowledge. Thus, in this scheme, there can be no question of God's role in Arnauld's freely choosing

[13] Ibid., 1603. [14] Grua, 345.

to be a priest rather than a manager of a hedge fund. In Leibniz's scheme, if Amauld exists, then he is no hedge fund manager, no matter how God may feel about that. There simply is no space for the scholastic debates previously noted to take hold. Furthermore, I think that when we fully elaborate the various replacement analyses that Leibniz proposed for counterfactuals, an element requiring post-volitional knowledge on God's part surfaces in every case. I made an effort to formulate the point involved in 'Leibniz on Divine Foreknowledge'.[15] Michael Griffin, while agreeing with the post-volitional aspect of the analysis I offered, criticized the details, and offered an alternative in 'Leibniz on God's Knowledge of Counterfactuals'.[16] His thesis that bears on our present concerns is summarized in his article as follows (p. 332): 'God knows a counterfactual by knowing what happens in the best possible world in which the antecedent is true, because he knows that that's the world he would have chosen to admit to existence, were he to choose among them.' Griffin is aware that, in virtue of superintrinsicalness, it cannot be the very antecedent of the relevant counterfactual that is true in some possible world that fails to obtain, if that antecedent refers to an existent individual substance. Leibniz's various efforts at constructing a replacement analysis offer different candidates for the proposition that is to replace the relevant conditional's antecedent. But in each case God's knowledge of the conditional offered as a replacement appears to require knowledge of what He would have willed in various counterfactual circumstances. Griffin improves on my account by noting that, in each relevant case, God need only heed a single divine decree, His primary free decree, to do the best possible. Griffin summarizes the point as follows:

> Through his intellect he [God] knows what happens in various possible worlds and what relations of similarity and relative optimality hold among these worlds. These things lie outside the scope of his will. But, his knowledge of counterfactuals also depends on his 'primary free decree' to always act in the most perfect way (DM 13), which determines the choices he would make in counterfactual circumstances.[17]

The plot grows thick at this point. As will become evident in the next section of this paper, I believe that Leibniz's mature, settled view is that God's 'primary free decree'—to act always in the most perfect way—is a decree that God elicits necessarily. That is, I believe that Leibniz

[15] *Faith and Philosophy*, 11 (1994), 547–71. [16] *Philosophical Review*, 108 (1999), 317–43.
[17] Griffin, 'Leibniz on God's Knowledge of Counterfactuals', 332.

held that the proposition—God always acts in the most perfect way—is metaphysically necessary, where 'metaphysically necessary' is used in the sense in which we ordinarily use that expression. This thickens the plot because it raises a question about where we are to locate God's knowledge concerning His 'primary free decree'. The ground for regarding it as pre-volitional is that that is where His knowledge of necessary truths gets located in the relevant tradition. And the ground for regarding it as post-volitional is that it is knowledge of a volition, and indeed, the fundamental volition whereby the world comes to be actual, and not just possible.

What is not murky in this matter is this: Leibniz's scheme for characterizing God's knowledge, based on peculiarities of his metaphysics, differs in substantive ways from that of the Scholastics that concern us, making inferences concerning his account of free choice, based on a comparison of his account of divine knowledge with Scholastic accounts, dubious.

There are other, important considerations that Michael Murray offers in support of attributing something quite close to the moral necessitarians' account of freedom to Leibniz. Some of these are considered in the next, and last, section of this paper.

IV

Those who claim that Leibniz's account of human freedom relies on the notion of moral necessity, as developed by the moral necessitarians, point to Leibniz's use of the expressions, actually the Latin and French for, 'moral necessity' and 'inclines without necessitating' and, in particular, the similarity in the way moral necessitarians and Leibniz use these expressions to solve similar problems. Consider the latter expression, 'inclines without necessitating'. The moral necessitarians used it in something like the following way: the obtaining of state of affairs X inclines toward, but does not necessitate, the obtaining of state of affairs Y, just in case the obtaining of X counterfactually implies the obtaining of Y, but the obtaining of X neither metaphysically nor causally necessitates the obtaining of Y. And the idea is that for any free choice C elicited by any human agent A there are circumstances R that obtain that incline, but do not necessitate A to elicit C. Moreover, the idea is that this relation of inclining, but not necessitating, is unique in the created world to the free choices of creatures.

It is the moral necessitarian's way of attempting to capture the idea that, when placed in suitable circumstances with appropriate powers and an appropriate patient to operate on, non-rational agents act in a way that is, as they say, determined to one outcome—that is, causally necessitated. Not so, us and angels.

I find this latter idea, that the relation of inclining, but not necessitating, is not exemplified in the case of the actions of non-rational agents, lacking in Leibniz. And this suggests to me that Leibniz had a quite different conception of what was meant by the expression 'inclines but does not necessitate' from the moral necessitarians. Nothing is ever simple in these matters; I don't expect the text I am about to cite to settle matters. But here is the sort of text that moves me. Attempting to sort out the notions of determination and necessity that occur in Locke's account of freedom, Leibniz wrote (in *New Essays*):

> There is no less connection or determination among thoughts than among motions (since being determined is not at all the same as being pushed in a constraining way) . . . It must be admitted that when one thing follows from another in the contingent realm the kind of determining that is involved is not the same as when one thing follows from another in the realm of the necessary. Geometrical and metaphysical implications necessitate, but physical and moral ones incline without necessitating.[18]

And the context makes clear that by the physical here Leibniz means physical, i.e., he was talking about a spatial movement of a non-rational being brought about by the spatial movements of other non-rational beings. The corresponding notion of physical necessity is not the one that predominates in Leibniz's writings, where usually physical necessity refers to what takes place in virtue of the natures of created things, whether they are material or immaterial. So here is my understanding of what Leibniz had in mind when he talked about 'inclining without necessitating': the obtaining of state of affairs X inclines toward, but does not necessitate, the obtaining of state of affairs Y, just in case it is causally but not metaphysically necessary that if X obtains then so does Y. With this as background and as an indication of the kind of account to follow, I am ready to display my unsophisticated, unnuanced, traditional account of Leibniz's thinking

[18] A/6/6/178.

concerning human free choice, and, in particular, the role in it of the notion of moral necessity.

In numerous texts Leibniz stated that while metaphysical necessity is inconsistent with freedom of choice moral necessity is not. Surprising as it may seem, the notion of metaphysical necessity involved in this claim is not nearly so transparent as one might have hoped, but it is not our present concern.[19] However complicated it may be, things get no easier when we turn to the notion of necessity that Leibniz claimed to be compatible with freedom of choice—namely, moral necessity. Consider the following passage from 'Causa Dei', a Latin summary of the main philosophical points of the *Theodicy*: 'Freedom excludes metaphysical necessity, the opposite of which is the impossible, i.e., what implies a contradiction. However, it does not exclude moral necessity, the opposite of which is the unfitting (*inconveniens*).'[20] The notion of moral necessity employed here is purely deontic, equivalent to moral obligation. But Leibniz did use a notion of moral necessity in a quite different sense to explain actions, including choices. And, in fact, I think Leibniz really needed such a notion in the passage just quoted; what he wrote there just does not make much sense except by trading in the deontic notion employed for something else. Consider the following passage from the *Theodicy*, where Leibniz was discussing moral necessity: 'it is necessary that the blessed should not sin, that the devils and the damned should sin, and that God should choose the best' (T #282). Here we have a non-deontic notion of moral necessity. Leibniz did not intend to affirm that it is morally obligatory that the damned should sin; one suspects that his view was that it is morally obligatory that they not sin, which, given their behavior, is why they are damned. We need to say something about Leibniz's use of the non-deontic notion of moral necessity. I take passages like the following from Leibniz's Fifth Letter to Clarke as canonical: 'All the natural forces of bodies are subject to mechanical laws, and all natural powers of spirits are subject to moral laws. The former follow the order of efficient causes; and the latter follow the order of final causes' (#124).

By my lights, what are here termed final causes are distinguished from efficient causes by the fact that they pertain to the causation of mental

[19] See the last paragraph of this paper for an example of the relevant complexity.
[20] Leibniz, 'Causa Dei', # 21.

phenomena. Not all final causes generate moral necessity in Leibniz's scheme of things, not even in the case of human choices. I think that the basic idea Leibniz had in mind is something like this: 'Choice C elicited by agent A is morally necessary, just in case; (1) the beliefs, motives, perceptions, desires, etc., of A causally necessitate that A elicits C; and (2) aspects of A's moral character play a causal role in the causation of A's eliciting C.'

Leibniz appears to utilize three varieties of necessity—metaphysical, physical, and moral; four varieties of laws—mechanical laws, laws of efficient causation, moral laws, and laws of final causation; and two varieties of causation—efficient and final. Consider the following scheme of simplification, which, I admit, is somewhat too tidy to fit all the relevant texts. Assume that mechanical laws are laws of efficient causation, and that moral laws—now taken non-deontically—are laws of final causation; employ the notion of a law of efficient causation in the obvious way to characterize physical necessity, and in an analogous fashion employ the notion of a law of final causation to characterize moral necessity. In conjunction with this I recommend the following thesis: for Leibniz, there is a univocal notion of causation involved in efficient and final causation, at least as the latter applies to choices of human agents; they differ simply in this—efficient causation holds among material entities, whereas final causation is causation among immaterial entities. I recommend an analogous thesis with respect to the relation of physical and moral necessity in Leibniz's mature scheme—there is a univocal notion involved; they differ simply in their domains. Physical necessity arises from the laws of efficient causation; moral necessity, from the laws of final causation. Moreover, I recommend the claim that the univocal notion of necessity involved is our notion of causal necessity—supposing there is such a thing.

No doubt there is some oversimplification involved in this scheme. I think there are loose threads in Leibniz's use of the notion of moral necessity. But I am not yet convinced by those who see the threads bound together by Leibniz's acceptance of the main theses of the moral necessitarians. Additionally, I am aware that quite special difficulties arise in trying to explicate Leibniz's use of the notion of moral necessity to characterize God's choices, and, in particular, God's choice to create the best possible world. That's another story. I took it on in a review of what I called in the review 'The Best Study of Leibniz's Philosophy Available'—Robert

Adams's book, *Leibniz: Determinist, Theist, Idealist*.[21] My basic idea expressed therein is this: When Leibniz claimed in the *Theodicy*, for example, that it is morally necessary, but not metaphysically necessary, that God chose to create the best possible world, the negative aspect of his claim,—that is, the claim that it is not metaphysically necessary that God chose to create the best possible world—can be captured in the following proposition: God chose among alternative possible worlds, each of which is possible in its own nature. And I claimed that this latter proposition is consistent with the claim that the proposition—God chose to create the best possible world—is metaphysically necessary in the usual sense of that expression. It may well be that, once again, I have oversimplified matters. Perhaps so. But I do not yet see that appeal to the concepts and theses of the moral necessitarians is what we need to get things just right.

[21] R. C. Sleigh, Jr., 'Leibniz on Freedom and Necessity', *Philosophical Review* (1999), 245–77.

7

Leibniz on Final Causation

MARLEEN ROZEMOND

One of the fundamental planks of early modern philosophy was its rejection of Aristotelianism. Prominent in this rejection was an abandonment of final causation in favor of efficient causation. But Leibniz thought differently. On a number of occasions he presented final causes as having a role equal in prominence to the role of efficient causes. For instance, in the *Monadology* he wrote:

Souls act according to the laws of final causes, through appetitions, ends and means. Bodies act according to the laws of efficient causes or of motions. And these two realms, that of efficient causes and that of final causes are in harmony with each other.[1]

And in the *Principles of Nature and Grace*:

And the perceptions in the monad arise from one another by the laws of appetites, or by the laws of final causes of good and bad that consist in notable perceptions, ordered or disordered. Similarly the changes in bodies and external phenomena arise from one another by the laws of efficient causes, that is, of motions. Thus there is a perfect harmony between the perceptions of the monad and the motions of bodies, which is first pre-established between the system of efficient causes and the system of final causes. And in this consists the accord and union between soul and body without one being able to change the laws of the other.[2]

[1] M 79. References to Leibniz should be understood as follows. (Dut.) stands for L. Dutens (ed.), Leibniz, *Opera Omnia* (6 vols.; Geneva, 1768); (M) for *Monadology*; (PNG) for *Principles of Nature and Grace*; (NE) for *New Essays*. References to Leibniz's work in the original languages can mostly be found in *Die Philosophischen Schriften von Gottfried Wilhelm*, Leibniz, ed. C. I. Gerhardt, 7 vols. (Berlin: Wiedmann, 1875–90; repr. Hildesheim: Georg Olms, 1978) (G). Translations can be found in *Philosophical Essays*, ed. Roger Ariew and Daniel Garber (Indianapolis, Ind: Hackett, 1989) (AG); *Philosophical Papers and Letters*, ed. Leroy E. Loemker (Dordrecht: Reidel, 1969) (L); and *Leibniz's 'New System' and Associated Contemporary Texts*, ed. and trans. R. S. Woolhouse and Richard Francks (Oxford: Clarendon Press, 1997) (WF).

[2] PNG 3; see also Fifth Letter to Clarke, G, viii. 419/L, 716–17.

These remarks are striking for more than one reason. First, Leibniz clearly departs from his early modern context in embracing final causes. Moreover, given that for him monads are more fundamental than bodies, these passages suggest that for him final causes are more fundamental than efficient causes. Finally, these remarks are puzzling also from an Aristotelian point of view: Aristotle did not see final and efficient causes as types of explanation that apply to different explananda; rather they are different but connected aspects of a single full explanation. To separate the two and allow for one to operate without the other is very puzzling from this perspective. So how should we understand Leibniz's use of the notion of final causation?

The impression that Leibniz separated final and efficient causes and confined each to a separate realm is just that, an impression. We will see that on other occasions Leibniz assigned both types of causality to each realm. But we will also see that Leibniz did really depart from other early moderns in the importance he attached to final causes. Leibniz's respect for final causation is in line with his repeated claim that certain aspects of scholasticism are in fact more useful than many of his contemporaries had acknowledged. But the connections with the Aristotelian scholastic background are more complex than one might think. First, early modern criticisms of final causation are central to our conception of the period, but what is less well known (among early modern scholars) is that final causation had already troubled the Aristotelian scholastics for centuries, and the worries can be traced back as far as Avicenna. Various scholastics had argued that final causation requires knowledge that only an intelligent agent could have. At the same time Descartes went beyond the scholastics in arguing that 'immanent teleology', the internal directedness at an end the Aristotelians admitted even for non-intelligent, natural agents, requires cognition on the part of an agent.[3] Second, I will relate Leibniz's use of the notion of final causation to his revival of the notion of substantial form, which he said should be understood on the model of the self, that is, a mind. Oversimplifying a bit, the notion of substantial form turns into the

[3] I take the term 'immanent teleology' from Margaret Osler. She argues against the understanding of early modern mechanists as systematically rejecting final causation. See her 'From Immanent Natures to Nature as Artifice: The Reinterpretation of Final Causes in Seventeenth-Century Natural Philosophy', *Monist*, 79 (1996), 388–408, and 'Whose Ends? Teleology in Early Modern Natural Philosophy', *Osiris*, 16 (2001), 151–68.

notion of a monad, the mind-like entities he regarded as the fundamental constituents of reality.[4] I will argue that for Leibniz genuine causal power requires final causality and cognition in the agent. Part of my argument is that Leibniz adopts Aristotelian ideas, final causation and substantial form, through a Cartesian lens.

I will begin with a discussion of final causation in the scholastics and Descartes's criticism of immanent teleology (section 1). Next I will relate Leibniz's revival of final causation to his resurrection of the notion of substantial form and argue that he saw final causation as connected to cognition (section 2). Then I will turn to the question of the apparent exclusion of efficient causation from the realm of monads (section 3).

1. Final Causation before Leibniz

For Aristotle, an explanation of a change involves appeal to four types of causes: final, efficient, formal, and material. In Aristotelianism, final causation was quite prominent, indeed, efficient causation was regarded as subordinate to it. In Aristotelianism, the orderliness and regularity of nature was due to the ends of nature, not to laws of efficient causality. An efficient cause acts in view of an end or goal, which is the final cause. This is easy to see in the case of an artisan producing an artifact: in view of the goal of making a statue, the artisan exercises efficient causality in such a way that she realizes her goal. For an Aristotelian this happens in nature as well: natural (non-intelligent) agents have specific powers to exercise efficient

[4] In the literature on Leibniz there has been intense controversy over the last two decades concerning the question whether during his middle years, roughly 1684–1704, Leibniz accepted the reality of corporeal substances of an Aristotelian type: composites of matter and substantial form. If so, for this period the familiar idealist interpretation according to which only monads are fundamentally real would not be accurate. The discussion was ignited by Daniel Garber, who favors the Aristotelian interpretation. See his 'Leibniz and the Foundations of Physics: The Middle Years', in Kathleen Okruhlik and J. R. Brown, *The Natural Philosophy of Leibniz* (Dordrecht: Reidel, 1985), 27–130. For defenses of the idealist interpretation, see, for instance, Robert M. Adams, *Leibniz: Determinist, Theist, Idealist* (New York: Oxford University Press, 1994); and R. C. Sleigh, Jr., *Leibniz and Arnauld: A Commentary on Their Correspondence* (New Haven, Conn.: Yale University Press, 1990). More recently, some interpreters have argued that Leibniz was a realist about corporeal substances even in his later years (Glenn A. Hartz, *Leibniz's Final System* (London and New York: Routledge, 2007); Pauline Phemister, *Leibniz and the Natural Word: Activity, Passivity and Corporeal Substances in Leibniz's Philosophy* (Dordrecht: Springer, 2005.) My own sympathies lie with the idealist interpretation, but I take my overall argument in this paper to be compatible with (versions of) either.

causality in view of their ends. In Aristotelian scholasticism, this idea took the form of saying that God had certain ends in view of which he gave creatures powers of efficient causality so that they can serve these ends.

But the scholastics struggled with final causality, and the worries go back at least as far as Avicenna. The late scholastic Francisco Suárez (1548–1617) wrote that 'although the final cause is in some sense the most important one, and is also prior to the other [types of causes] its causality [*ratio causandi*] is also more obscure [than the causality of the other types of causes]'. (dm xxiii.1).[5] The following discussion is heavily indebted to Dennis Des Chene's *Physiologia* and Anneliese Maier's 'Das Problem der Finalkausalität um 1320'.[6] I will pay special attention to Francisco Suárez. Suárez is particularly useful for understanding the scholastic background to the early moderns, since he systematically summarizes earlier discussions, he was very influential in the early modern period, and was sometimes cited by Leibniz.

An important question about final causation was how an end can exercise causality since it often does not (yet) exist. From early on philosophers espoused the idea that final causation requires knowledge of the end by an intelligent agent. This gave rise to the idea that it is the end as known by an intelligent agent, that is, the representation of the end by the agent, that exercises final causality, a view adopted by Avicenna. Averroes objected, however, that the end as known is not what one aims for, rather the form in reality is what one aims to achieve.[7] I aim for a really existing paper on Leibniz, not merely a paper that exists in my thought.

As Suárez's discussion illustrates, the central case of final causation was the case of created intelligent agents, and, he writes, the best-known case is ours, so he focuses on that case.[8] Suárez raises the question whether final causation applies beyond the case of created intelligent agents to God, on one hand, and to non-intelligent agents, which have no knowledge of ends,

[5] For the references to Suárez, see his *Disputationes metaphysicae* (DM) in *Opera Omnia*, ed. Charles Berton xxv–xxvi (Paris: Vivès, 1866; repr. Hildesheim: Georg Olms Verlag, 1998), referred to by disputation, section, and article, and his *De anima*, referred to by book, chapter, and section, to be found in *Opera Omnia*, vol. iii.

[6] Des Chene, *Physiologia: Natural Philosophy in Late Aristotelian and Cartesian Thought* (Ithaca, NY: Cornell University Press, 1996), 168–211; Anneliese Maier, 'Das Problem der Finalkausalität um 1320', in *Metaphysische Hintergründe der spätscholastischen Naturphilosophie* (Rome: Storia e letteratura, 1955), 273–335.

[7] Maier, 'Das Problem der Finalkausalität um 1320', 282.

[8] DM, xxiii. 1.8.

on the other hand. For our concerns the case of natural, i.e. non-intelligent, agents is the one that matters.[9]

As was common among the scholastics, Suárez included final causes in the explanation of natural phenomena (as opposed to the actions of created intelligent agents) in virtue of God's plans. Suárez relies on a widely used analogy with the role of an archer in making an arrow go for its target: 'natural agents are not so much said to act for an end, as being directed to an end by a superior agent',[10] and

> There is no proper final causation in actions insofar as they come from natural agents, but only a tendency [*habitudo*] to a certain endpoint [*terminus*], but insofar as they come from God there is final causality in them, insofar as there is in other external and *transeunt* actions of God. For the adequate principle of these actions is not only the proximate natural agent, except insofar as *secundum quid* namely in such an order; but the absolute principle is the first cause; therefore the adequate principle of such actions includes the intellectual cause intending their end.[11]

So, since for Suárez final causality requires an intelligent agent, in the case of natural, non-intelligent agents the final cause of the effects they produce does not lie simply in these agents but includes God's intentions in virtue of which ends exercise final causality.[12]

[9] Buridan's discussion of final causation deserves special mention. Endorsing the claim that final causation requires knowledge by an intelligent agent he argued that this means that it is the mental state of the agent that really acts as a cause. For Aristotle, when we ask why someone performs a certain action, the answer lies in the effect the agent aims to achieve, and that end is the final cause. According to Buridan, the answer to this question is 'the intention or volition or causes that are prior in being'. 'When it is asked: ''On account of what cause [*propter quod causam*] do you go to church?'', it must be said that it is because I intend or I want to hear the mass, and ''why does the doctor give medicine?'' the answer is: ''because he wants to heal.''' So it is a mental state in the agent that is the explanation, the cause, and Buridan argues that this mental state is the *efficient*, not the final cause (Maier, 'Das Problem der Finalkausalität um 1320', 310 ff.). In agents other than created intelligent agents, there is no genuine final causality and the orderliness of nature is not due to final causes, according to Buridan. Thus, Maier argues, Buridan eliminated final causality in favor of efficient causality in natural philosophy (ibid., 334–5; see also Des Chene, *Physiologia*, 186–7). Buridan's views resonate in Suárez's discussion of final causation. Suárez argues that final causation is a genuine type of causation, but the list of objection to this view that he offers overlaps substantially with Buridan's (DM, xxiii. 1.1–6; Maier, 'Das Problem der Finalkansalität um 1320', 301).

[10] DM, xxiii. 10.5.

[11] Ibid., 10.6.

[12] Des Chene writes that while there was a trend, starting with Ockham, to limit the application of final causation to intelligent agents, in the very late scholastics Des Chene discusses (sixteenth–seventeenth centuries) there was a return to a broader application of final causation that he also attributes to Aquinas (Des Chene, *Physiologia*, 169). He does not explain exactly how Aquinas's application was broader. For Aquinas, too, teleology requires intelligence, in natural agents

Suárez also makes clear the intimate connection between final and efficient causes in this interesting passage. Final causation accounts for the orderliness of nature, its regular natural behavior, Suárez explains: 'in virtue of its natural motion a stone is carried down, fire always heats, from different kinds of seeds different living beings are produced'.[13] Various natural properties flow from (*dimanatio*) the substantial form of a natural being, from its substantial form, as a result of 'an efficient cause which is subordinate to a final cause'.[14] For Suárez, creatures produce these characteristic operations through efficient causality, and they have been given the powers to produce these operations in view of certain ends intended by God.

The requirement of intelligence rather than mere cognition within scholasticism is striking. As we shall see in a moment, Descartes sometimes charged that end-directedness in an agent implies that the agent have cognition, but he did not make the stronger charge that it implies that the agent has intelligence. Why did the scholastics require intelligence? What about animals, can't they act in view of ends simply by knowing these ends even if they are not intelligent? One reason seems to lie in the following consideration: animals may have cognition of something that is an end of their action, but they cannot see an end as an end, and they cannot judge that something is good: full-blown final causality requires both.[15] Suárez noted that animals always use the same means towards an end, and this means they cannot exercise full-blown final causality.[16] No doubt he thought it means they did not deliberate, but act by natural necessity.

So the teleological nature of this picture consists in two stages for natural agents: (1) full-blow teleology requires cognition of ends by an intelligent agent, and God fulfils this role; (2) God places powers to achieve his ends in the natural agents. That means that these agents are endowed with immanent teleology: they are internally directed at ends. Indeed, for Aquinas and others efficient causation essentially involves such internal

divine intelligence does the job, as is clear from his argument from the occurrence of final causation in nature to God's existence. See *Summa Theologiae* (ST), I, qu. 2, art. 3: 'Beings that lack knowledge cannot tend towards [*tendunt in*] an end, unless directed by some knowing and intelligent being, as the arrow is directed by the archer. Therefore some intelligent being exists by whom all natural things are directed to their end and this being we call God' (New York: Blackfriars and McGraw-Hill, 1969).

[13] DM, xxiii. 10.3. [14] Ibid., 11.7. [15] Des Chene, *Physiologia*, 194–202.
[16] DM, xxiii. 10.12.

directedness: it is crucial that a causal power is a power to do something in particular, and thus the final cause is essential to the efficient cause:

> The efficient cause is the cause of the final cause inasmuch as it makes the final cause be, because by causing motion the efficient cause brings about the final cause. But the final cause is the cause of the efficient cause, not in the sense that it makes it be, but inasmuch as it is the reason for the causality of the efficient cause. For an efficient cause is a cause inasmuch as it acts, and acts only because of the final cause. Hence the efficient cause derives its causality from the final cause.[17]

> Now the movement of every agent tends to something determinate: since it is not from any power that any action proceeds, but heating proceeds from heat, and cooling from cold; wherefore actions are differentiated by their active principles. Action sometimes terminates in something made, for instance building terminates in a house, healing ends in health: while sometimes it does not so terminate, for instance, understanding and sensation. And if action terminates in something made, the movement of the agent tends by that action towards that thing made: while if it does not terminate in something made, the movement of the agent tends to the action itself. It follows therefore that every agent intends an end while acting, which end is sometimes the action itself, sometimes a thing made by the action.[18]

So an efficient causal power is inherently directed at something, an idea that is fundamental to the understanding of natural change as transitions from potency to act.

Descartes, however, thought that natural agents having such directed powers requires that they have cognition. In the Sixth Replies he explained this point in particular for the notion of heaviness, the tendency a body has to go down. This notion, he contends, is taken from the idea of the mind in several respects. The relevant part for our purposes is this:

> But what makes it especially clear that my idea of heaviness was taken partly from the idea I had of the mind is the fact that I thought it carried bodies towards the center of the earth, as if it had some cognition of it within itself. For this surely could not happen without knowledge, and there can be no knowledge except in a mind.[19]

[17] Aquinas, *Commentary on Aristotle's Metaphysics*, n. 775.

[18] Ibid., *Summa contra gentiles*, 3, Q2.

[19] I use the standard references to Descartes's work by volume and page number. For the texts in the original languages, see *Œuvres de Descartes*, ed. Charles Adam and Paul Tannery, 12 vols. (Paris: Vrin, 1964–74) (AT), vii. 441–2. For translations see *The Philosophical Writings of Descartes*, trans.

So Descartes claims that attributing heaviness to bodies means ascribing knowledge to them, knowledge of where they are supposed to go. The heaviness in virtue of which bodies have a tendency to go down according to the Aristotelians is thus an anthropomorphic quality, in Descartes's view.[20] He does not talk about final causation explicitly here, but he is taking on an understanding of heaviness as a quality that operates by directing a body at an end, a place where it is supposed to go.[21]

In sum, the scholastic and Cartesian background strongly suggest that final causation requires mentality: full-blown teleology requires intelligence according to the scholastics. Furthermore, the immanent teleology they attribute to natural, non-intelligent agents requires cognition according to Descartes. Finally, this immanent teleology is required for efficient causality on a scholastic understanding of such causality. On this understanding an agent has efficient causality in virtue of powers to achieve specific ends. Putting these points together: an agent exercising efficient causality implies cognition.

2. Final Causes and Cognition in Leibniz

In the realm of bodies Leibniz's views relate to this historical background in the first place in a fairly straightforward manner. Descartes had argued that we should not discuss final causation in natural philosophy because doing so involves investigating God's purposes and these are unknown to us (*Principles of Philosophy* I.28).[22] Leibniz agreed with Descartes that

John Cottingham, Robert Stoothoff, and Dugald Murdoch, 3 vols. (Cambridge: Cambridge University Press, 1985–91) (CSM), ii. 297–8. CSM provides the AT page numbers in the margins.

[20] For discussion of the relation between Descartes's criticism of teleological explanations and Aristotelian scholastic practice, see Des Chene, *Physiologia*, esp., 168–71 and 391–8.

[21] Similar ideas can be found in Boyle. See *The Works of Robert Boyle*, ed. Michael Hunter and Edward B. Davis, 14 vols. (London: Pickering and Chatto, 1999–2000), xi. 110. I owe this reference to Lawrence Carlin; for discussion, see his two unpublished papers on Boyle: 'Final Causes in Robert Boyle: The Question of Immanent Finality' and 'Teleology and Systematization in Boyle's Natural Philosophy'.

[22] *Principles of Philosophy*, i. 28. Descartes does rely on knowledge of God's nature, but not God's purposes in deriving the laws of nature, because he derives the fundamental laws of motion from God's immutability (see *ibid.*, ii. 37–42). So the disagreement between Descartes and Leibniz does not merely lie in a disagreement about epistemic access to God's purposes, since for Descartes God's purposes are not relevant to the laws of motion. For discussion of Descartes and Leibniz on the laws of nature, see in particular Garber, 'Mind, Body, and the Laws of Nature', *Midwest Studies in Philosophy* 8 (1983), 105–33 Leibniz was certainly not the only early modern who disagrees with Descartes about the question

bodily phenomena can, and indeed, should, be explained in terms of the laws of efficient mechanical causation. But the origin of these laws, he claims repeatedly, lies in final causes: they were chosen by God in view of his purposes, and so they are subordinate to God's ends.[23] And Leibniz rejected Descartes's claim of the epistemic inaccessibility of divine purposes, and agreed with, for instance, Boyle on this issue. Indeed, he argued that considering purposes can actually be helpful in determining what the mechanical laws of nature are. For instance, he claimed that Snell had discovered his law of refraction by considering final causes.[24]

So Leibniz does relate final causes to the realm of bodies, although their role is indirect; God's purposes explain what mechanical laws obtain, but explanations in the bodily realm run in mechanistic terms. And Leibniz's view bears a clear similarity to the scholastic picture, on which natural agents get their powers to produce and tendencies to certain ends from God in view of his purposes. For Leibniz, mechanical laws are chosen by God in view of his purposes. This similarity between the two views is a bit superficial, however, because it does not yet address the Cartesian criticism of immanent teleology. The scholastic picture included immanent teleology in natural agents, for Descartes immanent teleology implies knowledge on the part of the agent and so he denied immanent teleology in bodily agents. What is Leibniz's stance on this issue?

This question is answered, I believe, in the course of Leibniz's criticism of Descartes's conception of material substance as essentially extended and utterly passive. Leibniz argued that this conception of material substance is unsatisfactory, and needs to be supplemented with a notion of force. I will argue that the way Leibniz develops this criticism means that he accepted the Aristotelian idea that a genuine efficiently causal power is directed at ends, and that he accepted Descartes's claim that internal directedness at ends implies knowledge on the part of the agent. But, unlike Descartes, Leibniz thought that we need to accept such powers, forces, both to explain bodily occurrences and in view of the requirements for substancehood. He thinks this means we have to go beyond the strictly material and appeal to substantial forms or monads, which are cognizing entities.

whether we should investigate God's purposes in nature. Boyle, for instance, contends that we must do so (*Works*, xi. 81). A failure to do so could lead to a 'loss of benefits relating to Philosophy as well as Piety'.

[23] *Discourse on Metaphysics*, 19–22 (NE, 179); *Specimen dynamicum* (AG, 126); Draft of *New System* (9, iv. 472/WF, 22).

[24] *Discourse on Metaphysics*, 22.

This interpretation requires that perceptions are cognitions, and monads genuine mental beings. Some interpreters have questioned this view. For instance, according to Robert McRae, in the absence of consciousness, perception does not count as cognition. John Carriero argues for an interpretation of Leibniz's notion of substance that emphasizes the importance of *activity* for this notion as opposed to mentality. He contends that teleology is crucial, but not cognition, and thinks that for Leibniz the final causality of monads does not involve cognition.[25] But, in my view, for Leibniz activity and mentality are connected. In this section I will first defend the view that Leibniz did see perceptions and consequently monads as mental, then I will return to the question of force and its connection with cognition. One way to recognize that Leibniz saw perceptions as mental is by focusing on the fact that monads are modeled on the human soul or mind. An additional reason derives from Leibniz's connecting perception to simplicity.

[25] Robert McRae, *Leibniz: Perception, Apperception and Thought* (Toronto: University of Toronto Press, 1976), 24; John Carriero, 'Substance and Ends in Leibniz', in Paul Hoffman, David Owen, and Gideon Yaffe, *Contemporary Perspectives on Early Modern Philosophy: Essays in Honor of Vere Chappell* (Guelph: Broadview Press, 2008). Margaret Wilson raises the question what it means for the states of monads to be perceptions given that they are not conscious ('Confused vs. Distinct Perception in Leibniz: Consciousness, Representation and God's Mind', in her *Ideas and Mechanism: Essays on Early Modern Philosophy* (Princeton, NJ: Princeton University Press, 1999).

In his repeated separation of the two realms of causation Leibniz usually speaks in terms of souls or minds. Minds have intelligence, and sometimes Leibniz offers a restricted use of the term 'soul', according to which a soul has sensation and memory, but not all monads do (M, 19) So one might think that the scope of final causation in these contexts is limited to only a part of the monadic realm. I do not agree, and at PNG, 3 Leibniz states the division of realms in terms of monads rather than souls. The statements in terms of souls should perhaps be read in view of Leibniz using the term 'soul' both in the strict sense noted above, but also in a broader sense where the term refers to all monads, a usage also noted at M, 19.

Robert McRae and Mark Kulstad suggest that Leibniz sometimes limited final causation to voluntary perceptions (McRae, 'Appetition in the Philosophy of Leibniz', 67; Kulstad, 'Appetition in the Philosophy of Leibniz' in Albert Heinekamp, Wolfgang Lenzen, and Martin Schneider (eds.), *Mathesis rationis: Festschrift für Heinrich Schepers* (Münster: Nodis Publikationen, 1990), 146). They refer to Leibniz's comments on Lamy's *De la connaissance de soi-même* where he writes: 'Without relying on the fact that the laws of motion are established in virtue of divine wisdom and are not at all geometrically necessary, it is sufficient to say that perceptions that express the laws of motion are just as connected as those laws, which they express according to the laws of efficient causes. But the order of voluntary perceptions is that of final causes, which conform to the nature of the will' (G, IV 580/WF, 155). McRae and Kulstad take the passage to say that the laws of final causes only apply to voluntary perceptions, and they take the passage to mean that the laws of efficient, mechanical causation apply (in some sense) to all others. I think the passage should not be taken in this way. Leibniz is responding to Lamy's concern about freedom, and so it is not surprising that he should limit himself to noting that voluntary perceptions fall under the order of final causes. Consequently the passage does not clearly have the implications McRae and Kulstad attribute to it.

Leibniz's notion of the monad evolved from his resurrection of the notion of substantial form, his most prominent and most central scholastic import. He argued that this notion was necessary to supplement the Cartesian notion of matter, which by itself he deemed unsatisfactory on both metaphysical and scientific grounds.[26] He explains some of the central metaphysical ideas of his critique in the following passage from the *New System:*

> I perceived that it is impossible to find the principles of a true unity in matter alone, or in what is only passive, since everything in it is only a collection or aggregation of parts to infinitely. Therefore in order to find these real entities I was forced to have recourse to a formal atom, since a material thing cannot be both material and, at the same time, perfectly indivisible, that is, endowed with a true unity. Hence it was necessary to restore, and is it were, to rehabilitate the substantial forms which are in such disrepute today, but in a way that would render them intelligible, and separate the use one should make of them from the abuse that has been made of them. I found then that their nature consists in force, and that from this there follows something analogous to sensation and appetite, so that we must conceive of them on the model of the notions we have of souls.[27]

Leibniz emphasizes here his view that we need recourse to something other than Cartesian matter because we need genuine unities, but he also mentions his other main reason for going beyond such matter: the need for something active. Indeed, he presents unity and activity as connected: he writes that something that is passive cannot have real unity. Both needs can be fulfilled by the notion of substantial form, he claims; a cleaned-up version of this notion provides us with an entity that can generate genuine unity by being indivisible and that is active.

From a historical perspective, Leibniz's use of the notion of substantial form does not immediately suggest he is talking about mental substances: on the contrary. In the Aristotelian tradition, at most some substantial forms are subjects of mental states: humans have substantial forms but so do animals, plants, and mixed bodies like gold and the elements. Leibniz is in line with this tradition when he argues that we need the notion of substantial form to generate corporeal substances as opposed to material

[26] I will leave the scientific issues aside. Leibniz argued that the Cartesian conception of matter with its focus on motion as opposed to force gets the laws of mechanics wrong. See, for instance, *Discourse on Metaphysics*, 17.

[27] G, iv. 478/AG, 139.

beings that are mere aggregates. But he departs from the Aristotelians when he explicitly models substantial forms on the human soul, as when he writes to Arnauld that we need 'a soul or substantial form on the model of what we call ''me'' [une âme ou forme substantielle à l'exemple de ce qu'on appelle moi]'.[28] And in the draft of the *New System*: 'what makes a corporeal substance must be something that corresponds to what is called ''me'' in us, what is indivisible and yet active'.[29] Furthermore, he focuses on features of the human soul that distinguish it from other substantial forms when he writes that the cleaned-up version of the notion of substantial form he wishes to use 'consists in force, and that from this there follows something analogous to sensation and appetite'.

What is more, it is worth noting that this approach means that from an Aristotelian scholastic perspective he uses a rather peculiar version of the notion of substantial form. From that perspective the human soul was an atypical, marginal type of substantial form. In the Aristotelian scholastic tradition, regular substantial forms are intrinsic constituents of substances—they can't exist separately—that is to say, they cannot exist without existing as a constituent of substances. The human soul is the only substantial form that has the capacity to exist apart, and this was important for the religious commitment to the survival of the soul after the death of the body. Aquinas and others defended its special status on the ground that the human intellectual soul has an activity that it performs without that action being an action of a bodily, ensouled organ.[30] Averroes had used the special nature of the human intellect to argue that the human intellectual soul cannot be the form of the body. Aquinas and others clearly felt a need to defend the possibility of such a substantial form.[31] These features of Aquinas's view of the human soul were often shared by seventeenth-century scholastics, including Suárez.[32]

The dispute with Averroes gave rise to the verdict by the Lateran Council of 1513 that philosophers should argue that the rational soul is the

[28] G, ii. 76/AG, 79. [29] G, iv. 473/WF, 23. [30] ST, I. 75.2.

[31] Aquinas addresses this anomalous feature of the human soul as substantial form, and argues that while the human soul can exist separately its *natural* place is in union with the body, just like a light body's natural place is up, even if it may happen to be down (ST I.76.1, ad 6).

[32] I discuss these issues also in my *Descartes's Dualism* (Cambridge, Mass.: Harvard University Press, 1998), chs. 2 and 5, and in relation to Leibniz in 'Leibniz on the Union of Body and Soul', *Archiv für Geschichte der Philosophie*, 79 (1997), 150–78.

form of the body, a decree Leibniz cites.[33] The Council was concerned not with the preservation of the traditional Aristotelian notion of substantial form but with the human soul's individuality and immortality. The verdict was issued in response to the Averroist view that the human intellect is not part of the human soul, of the form of the body. Averroes had inferred that there is only one intellect for all human beings and this posed a threat to individual immortality.[34] Descartes too cited this verdict by the Council, focusing explicitly on its demand that philosophers show the immortality of the human soul.

Leibniz's use of the human soul as the model for the substantial soul is surprising from an Aristotelian perspective, but it is not surprising in relation to Descartes, who sometimes labeled the human soul the only substantial form.[35] Leibniz sees himself as following Descartes in various ways on this issue, although he criticizes Descartes's restriction of substantial form to humans alone.[36] As Robert Adams argues, other early moderns also used the notion of substantial form in Descartes's way, and so Leibniz's use of a notion of substantial form where the human soul is its model amounts to an early modern interpretation of this notion.[37] In sum, by the time we get to Leibniz, the human soul has gone from marginal substantial form to the model of substantial form.

So now the question is this: *in what sense exactly* is the human soul the model for substantial form for Leibniz? He did not think that all substantial forms, and, later, all monads, are *exactly* like human souls: human souls, or minds, are special, because they have intelligence and free will, and not all monads have consciousness. When he explains in what sense the

[33] G, ii. 75/AG, 78.

[34] For the pronouncement by the Lateran Council, see Henrich Denzinger, *Enchiridion symbolorum, definitionum et declarationum de rebus fidei et morum* (Freiburg: Herder, 2005), 482–3, arts. 1440–1, and 390, art. 901. The issues that provoked these statements from the Lateran Council are discussed in detail in Étienne Gilson, 'Autour de Pomponazzi: problématique de l'immortalité de l'âme en Italie au début du XVI^e siècle'; and id., 'L'affaire de l'immortalité de l'âme à Venise au début du XVI^e siècle', in his *Humanisme et Renaissance* (Paris: Vrin, 1983). See also Eckhard Kessler, 'The Intellective Soul', in Charles B. Schmitt, Quentin Skinner, Eckhard Kessler, and Jill Kraye (eds.), *The Cambridge History of Renaissance Philosophy* (Cambridge: Cambridge University Press, 1988), 500–7.

[35] AT, vii. 356/CSM, ii. 246; AT, iii. 503, 505/CSM, iii 207–8; AT iv. 346/CSM, iii. 279. For discussion of Descartes's use of the notion of substantial form, see Paul Hoffman, 'The Unity of Descartes's Man', *Philosophical Review*, 95 (1986), 339–70; Rozemond, *Descartes's Dualism*, chs. 4–5.

[36] LA, 113; NE, 317–18; G, vi. 547.

[37] Robert Adams discusses in particular Boyle and Cudworth (Adams, *Leibniz: Determinist, Theist, Idealist*, 319–24).

human soul serves as a model, he sometimes argues that substantial forms in his system will have something *analogous to* sensation and appetite (*New System*, IV 479/AG 139, *On nature Itself*, G IV . . . AG 163).[38] In later texts he writes that all monads—the notion that evolves out of substantial forms—are characterized by perceptions and appetites *tout court*—without the qualification 'analogous'.[39] So the human soul's mental states are crucial to it being the model of the substantial form—in a relatively broad sense of mental that does not imply intelligence or consciousness.[40]

One might still hesitate to regard Leibnizian perceptions as mental, given that he denies that perception and appetite are always characterized by consciousness. On the other hand, the mere denial of consciousness does not obviously disqualify perception as cognition; Descartes has often been criticized for a failure to leave room for unconscious mental states. Indeed, Leibniz's own criticisms of Descartes on the ground that he failed to acknowledge unconscious perceptions would fall flat if (unconscious) perceptions were not in the end for him mental states. And it seems puzzling for him to speak of *perceptions* if he did not regard them as mental. My view is that perceptions are mental by being representational rather than conscious, but I will not defend this view here.[41] A full discussion of Leibniz's conception of the mental goes beyond the scope of this paper. I will now turn to the role of the notion of simplicity.

As we saw, one of the main reasons Leibniz adopted substantial forms was the need for entities that have genuine unity, which for him results

[38] In a letter to De Volder, Leibniz's statement suggests a possible bridge between the two types of phrasing. First, he writes: 'It is worthwhile to consider, however, that there is a maximum intelligibility in this principle of Action, because there is something in it analogous to what is in us, namely perception and appetite, since the nature of things is uniform and our nature cannot differ infinitely from the other simple substances of which the whole Universe consists' (G, ii. 270/L, 537). Now Leibniz does not qualify the application of the labels 'perception' and 'appetite' to monads other than human souls; he calls them perception and 'appetite' *tout court*, but labels them analogous to what is in us. Later in the passage he writes: 'Considering the matter accurately, moreover, it must be said that there is nothing in things except simple substances and in them perception and appetite.' So maybe Leibniz came to think one could call what exists in all monads perceptions and appetites without qualification, and so monads are analogous to our souls in this sense, but of course in his view not all are conscious or intellectual.

[39] M, 14; PNG, 3.

[40] This use of the term 'mental' deviates from Leibniz's own in so far as he reserved the term 'minds' for human souls.

[41] For a defense of the view that perceptions are representational for Leibniz, see Alison Simmons, 'Changing the Cartesian Mind: Leibniz on Sensation, Representation and Consciousness', *Philosophical Review*, 110 (2001), 31–75. Simmons spends little time explaining what representationality means. I think for Leibniz perceptions are intrinsically representational, that is, not merely in virtue of relations to the objects represented.

in a requirement of simple entities, which are the monads.[42] He describes perception as a type of expression, a prominent notion in his work. Body and soul express each other, for instance, and this example shows that expression is a term that does not connote the mental. But, as various commentators have noted, what makes perception a special type of expression is a connection with simplicity.[43] I will argue that given the historical context this connection with simplicity is a strong indication that he conceives of the perceiving monad as mental.

Leibniz explicitly connects simplicity and perception in a number of passages, quite prominently and repeatedly in the *Monadology*. There he defined perception in terms of its belonging to a simple substance: 'The passing state that contains [*envelope*] and represents a multitude in a unity or in a simple substance is nothing other than what one calls perception, which one must distinguish from apperception or consciousness as will become clear in what follows'.[44] We find the connection again in the well-known mill passage at *Monadology* 17, where Leibniz argues that perceptions cannot be explained mechanistically. He illustrates the point by asking us to imagine a thinking machine large enough so that you can walk into it, as into a mill. He claims that you would not find anything that explains perception, only mechanical states. Perceptions, he claims, requires a *simple* substance.[45]

In the historical context, this connection with simplicity strongly suggests that he thinks of substantial forms, monads, and perception as mental. In the Aristotelian tradition substantial forms were not generally simple, but the atypical human soul was. In late scholasticism the substantial forms of inanimate substances, and the souls of plants and lower animals, were supposed to be divisible. Human souls were regarded as indivisible, and there was controversy about the souls of the higher animals.[46] So indivisibility

[42] Some recent interpreters have suggested that not just monads (in the sense in which this notion is usually understood) but also corporeal substances, which include monads as their constituents, are simple and indivisible for Leibniz (see Hartz, *Leibniz's Final System*, 190–1; Phemister, *Leibniz and the Natural Word*, 74–5). I believe the response to this position requires a proper analysis of relevant types of simplicity. I cannot undertake to offer such an analysis here, however.

[43] See McRae, 'Appetition in the Philosophy of Leibniz', 24; Simmons 'Changing the Cartesian Mind', 42. Simmons offers some analysis of what the connection with simplicity means.

[44] *Monadology*, 14; see also ibid., 16.

[45] For discussion of this passage, and simplicity and perception more generally in Leibniz, see Marc Bobro and Paul Lodge, 'Stepping Back Inside Leibniz's Mill', *Monist*, 81 (1998), 554–73.

[46] For a useful discussion, see Des Chene, *Life's Form: Late Aristotelian Conceptions of the Soul* (Ithaca, NY: Cornell University Press, 2000), 171–89. The divisibility of the souls of plants and lower animals was illustrated by various phenomena: in the case of plants the fact that a cutting from a tree can live

was not a feature of substantial forms generally, but only of certain types of souls, most uncontroversially of the human soul.[47] So we see now a further sense in which for Leibniz the human soul was the model for the substantial form and the monad: its simplicity.

Furthermore, the connection between simplicity and the mental is even stronger outside the Aristotelian scholastic context. The idea that the human mind or soul is simple was widespread among non-Aristotelian early modern thinkers. Indeed, there is a rich history of arguments from the nature of the mental to the simplicity of the human soul and to its immateriality and immortality. Such arguments go back to Plotinus, and its central ideas have their roots in Plato, in particular his *Phaedo*, a work Leibniz cherished. The best-known discussion of a version of the argument occurs when Kant criticizes it in the Second Paralogism, while labeling it the 'Achilles of all dialectical inferences in the pure doctrine of the soul'.[48] According to this 'Achilles Argument', the unification and connection of mental contents requires a simple subject. In Kant's version the subject of thought must be simple, otherwise the parts of a thought would be scattered over the parts of the subject and nothing would think the entire thought. A version of the Achilles Argument occurs in a correspondence between Samuel Clarke and Anthony Collins, and provoked approval from Leibniz, who, indeed, himself had offered a version of the argument in an early work.[49]

and produce foliage, in the case of lower animals the example of a worm that continues to manifest life after being cut (Suárez, *De anima*, I.XIII, 2, 3).

[47] In the *New System* Leibniz writes that he remembers Aquinas saying that the souls of animals are indivisible. 'I saw that these forms and souls must be indivisible just like our mind, as in fact I remember was S. Thomas' opinion concerning the souls of beasts' (G, iv. 479/AG, 139). Earlier, in a letter to Arnauld, Leibniz had written that Aquinas said that substantial forms in general are indivisible; as Garber and Ariew note in their translation, this was probably not accurate. And now in the *New System* the focus is on a subset of substantial forms. But, on the other hand, Leibniz does not say that Aquinas held that the *human* soul is indivisible, it is the souls of animals. In a sense this fits the picture as I see it: Leibniz wants to be more generous than Descartes about substantial forms: humans are not the only ones who have them, and consciousness is not required.

[48] *Critique of Pure Reason*, A351.

[49] *Correspondance Leibniz–Clarke présentée d'après les manuscripts originaux des bibliothèques de Hanovre et de Londres*, ed. A. Robinet (Paris: Presses universitaires de France, 1957), 32. For Leibniz's early Achilles Argument, see his 'The Immortality of the Human Mind, Demonstrated in a Continuous Sorites', which is part of *The Confession of Nature against Atheists*, of 1669 (G, iv. 109–10/L, 113). For discussion of the history of the Achilles Argument, see Ben Lazare Mijuscovic, *The Achilles of Rational Arguments* (The Hague: Martinus Nijhoff, 1974); and T. Lennon and R. Stainton (eds.), *The Achilles of Rational Psychology* (Dordrecht: Springer, forthcoming). For Clarke's use of the argument, see my 'The Achilles Argument and the Nature of Matter in the Clarke–Collins Correspondence', ibid.; and 'Can

Relating Leibniz to this tradition is not an entirely simple matter, and this aspect of his thought involves both terminological and substantive differences with other early moderns. As we saw, unlike Descartes, Leibniz held that souls or substantial forms can be found not just in humans but everywhere in nature. He reserved the term 'mind' for human souls and while Descartes used the term 'thought' for the entire spectrum of mental states Leibniz tended to use the term 'thought' for an intellectual type of perception peculiar to minds. Another important difference lies in Leibniz's rejection of Descartes's view that all perceptions are conscious,[50] and he thought only some are intellectual.

It is tempting to thinking of the Achilles Argument as turning on the notion of consciousness, or on intellectual types of thought, so that it would be problematic to relate Leibniz's requirement of a simple subject for perception generally to the Achilles Argument. This would be a mistake. First, the argument pre-dates the early modern period and its notion of consciousness. Furthermore, statements of the argument in the period did not confine themselves to consciousness or intellectual thought. Clarke stated the argument in terms of consciousness and thought, but he made clear that he had a very broad notion of the mental in mind.[51] Pierre Bayle discusses a version of the Achilles Argument that focuses not on intellectual but on sensory states: 'For if a thinking substance was unified only in the way a sphere is, it would never see a whole tree at once; it would never feel the pain produced by the blow of a stick.'[52]

Matter Think? The Mind–Body Problem in the Clarke–Collins Correspondence', in Jon Miller (ed.), *Topics in Early Modern Philosophy of Mind* (Dordrecht: Springer, forthcoming). On a different version of this argument, the simplicity of the subject is not inferred from the connections between mental contents but from self-consciousness: the awareness of the subject of its own mental states. See Devin Henry, 'The Neoplatonic Achilles', in Lennon and Stainton, *The Achilles of Rational Psychology*.

[50] PNG, 4; M, 4.

[51] Clarke distinguishes his argument which focuses on 'bare Sense or Consciousness it self' from arguments that appeal to the higher capacities of the human mind: 'its noble Faculties, Capacities and Improvements, its large Comprehension and Memory; its Judgment, Power of Reasoning, and Moral Faculties' (see Samuel Clarke, *The Works* (London, 1738; repr. New York: Garland Publishing) (W), iii. 730). He offers a very specific definition of consciousness: '*Consciousness,* in the most strict and exact Sense of the Word, signifies neither a *Capacity of Thinking*, nor yet *Actual Thinking*, but the *Reflex Act by which I know that I think, and that my Thoughts and Actions are my own and not Another's*.' But at the same time he writes that in the context of the Achilles Argument this definition is not relevant 'because the Argument proves universally, that Matter is neither capable of this *Reflex Act*, nor of the first *Direct Act*, nor of the *Capacity of Thinking* at all' (W, iii. 784).

[52] Pierre Bayle, *Historical and Critical Dictionary: Selections*, trans. Richard H. Popkin (Indianapolis, Ind.: Hackett, 1991), q.v. 'Leucippus' (p. 130).

So Leibniz's view that perception is representation of a multiplicity in a unity, and that, indeed, it requires a simple being for its subject is part of a rich history of such views about the nature of the mental. This constitutes a very strong indication that he saw perception in general, and not just thought, as mental. In sum, there are strong reasons for interpreting Leibniz as conceiving of perception as mental.

We are now ready to turn to the relationship between efficient causality and mentality. In arguing against occasionalism Leibniz writes that we must admit that 'a certain efficacy has been placed in things, a form or force'.[53] So Leibniz links causal efficacy to force, and he explicitly connects force and perception. As we saw, he writes in the *New System* that the nature of substantial forms 'consists in force, and that from this follows something analogous to sensation and appetite; and hence we must conceive of them on the model of the notion that we have of souls'.[54] And we saw that in later related texts he leaves out the qualification 'analogous'. So Leibniz thinks force involves sensation and appetite, types of mental states.

Leibniz's recourse to substantial forms understood mentalistically makes sense in light of the Cartesian claim that immanent finality implies cognition. For Leibniz, genuine causal activity requires force, which he characterizes as a striving, *nisus, conatus, effort*, for an effect. So force implies immanent finality, and for this reason it implies perception and appetite.[55] This is why an appeal to force requires going beyond the physical to substantial forms modeled on the human soul.

Another aspect of the Cartesian background that helps explain this line of thought is the Cartesian conception of matter as utterly passive, which was grounded in the conception of the essence of matter as extension. Descartes himself is often thought not to ascribe any causal power to matter as a result. I do not agree with this interpretation, but it is clear that this view was often adopted by his successors. For Malebranche, creatures in general have no genuine causal power. The idea that matter is passive played a significant role in his arguments; matter is disqualified from causal efficacy by virtue of its very nature, which is passive, whereas mind is disqualified

[53] 'De ipse natura', G, 507/AG, 158. [54] G, iv. 479/AG, 139.

[55] Paul Lodge has suggested to me that this is not exactly right for the conatus of bodies. Their directedness, however, is parasitic on the final causality of the forces that constitute the nature of substances.

for other reasons.[56] Leibniz agrees that matter as extension is purely passive. He argues that we need to add the notion of force and, as a result of assuming that matter is purely passive, he conceives of force mentalistically. Thus the following picture emerges: for Leibniz genuine causal activity, force, is teleological and mental.

3. The Separation of Final and Efficient Causality

This leaves us with the question why Leibniz repeatedly suggests that the realm of monads is the realm of final causation only. If this were really Leibniz's view, then for him final causation would in fact be more fundamental than efficient causality, since monads are more fundamental than bodies, which are merely (well-founded) phenomena. Indeed, perhaps the only real causality is final causality: bodies are mere phenomena grounded in monads, and Leibniz holds that the laws of motion refer to forces, which ultimately are features of monads. Perhaps efficient causality is a notion one can use when speaking of the laws of nature, the regularities of the bodily world, but it does not refer to any type of *real causal power*, at least not a real causal power within the created world. That would be a very striking result. A version of this view has been defended by Sukjae Lee, who argues that there is no room for genuine efficient causality in creatures for Leibniz on the ground that all efficient causality resides in God.[57] But there is good reason to think that Leibniz did not exclude efficient causation from the realm of monads.

On two occasions, to my knowledge, Leibniz addresses the relationship between final and efficient causality, both are lesser-known texts. In *Specimen demonstrationum Catholicarum seu Apologia Fidei ex Ratione*, *Specimen of Catholic demonstrations of an Apology of the Faith from Reason* (dated *c.* 1685) he writes: 'I maintain that even final causes can be referred to efficient causes [*causas finales referri posse ad efficientes*], namely when the agent is intelligent, for then it is moved by the thought, and even moral causes are

[56] For discussion, see, for instance, Steven Nadler, 'Doctrines of Explanation', in M. R. Ayers and Daniel Garber (eds.), *The History of Seventeenth Century Philosophy* (Cambridge: Cambridge University Press, 1998), 536–42.

[57] Sukjae Lee, 'Leibniz on Divine Concurrence', *Philosophical Review*, 113 (2004), 203–48.

natural causes for they are of the nature of the mind'.[58] This is not a context in which Leibniz is focusing on monads, and, indeed, it is a text from the middle years where Leibniz's monadological views were not yet in full view. So we should not take Leibniz to claim that only intelligent monads as opposed to non-intelligent ones are subject to efficient causality. Rather, the context is one where he is addressing the view that final causes must not be attributed to nature and that they are not natural but made up by us. Since this text is not focused on the monadic level it is not as clearly useful for our purposes as the following passage, from the *Notes on Stahl*, dated 1704:

[T]he present state of body is born from the preceding state through the laws of efficient causes, the present state of the soul is born from its preceding state through the laws of final causes. The one is the place of the series of motion, the other of the series of appetites; the one is passed from cause to effect, the other from end to means. And in fact, it may be said that the representation of the end in the soul is the efficient cause of the representation in the same soul of the means [*et revera dici potest, repreaesentationem finis in anima causam efficientem esse repraesentationis mediorum in eadem*][59]

This text addresses our question head on: Leibniz starts by stating the separation of the two realms of causes, then adds that in fact efficient causes do apply in the realm of souls: he presents the representation of an end as an efficient cause.[60]

[58] Grua, 28.

[59] Dut. ii. 2.134. I owe this reference to Lawrence Carlin, 'Leibniz on Final Causes', *Journal of the History of Philosophy*, 44 (2006), 217–33.

[60] This is not to say that final causes *are* efficient causes. Leibniz does not *identify* final cause and efficient cause here: the final cause is the end itself; the efficient cause is the mental representation of the end. Carlin (ibid.) argues that for Leibniz final causes are a species of efficient causes. While my question was whether there is efficient causality at the level of monads, Carlin proceeds by asking whether final causes *are* (a species of) efficient causes. From an Aristotelian point of view, that is a surprising approach, given that final and efficient causes were different types of explanation, or rather different aspects of one full explanation. (Anneliese Maier does cite a less-known scholastic, Guido Terreni, as claiming that the activity of an end is not really different from that of an efficient cause. See Maier, 'Das Problem der Finalkausalität um 1320', 286). The texts Carlin cites are the ones I cite above. But in neither text does Leibniz say that final causes *are* efficient causes. Carlin's discussion is not always careful about the distinctions between the end itself, the end as represented, and the mental act in which the end is represented. I am not certain that Leibniz himself is always careful about this either, although in some of his remarks his point is precisely to draw such distinctions. In allowing a role for both types of causality within souls, Leibniz's view now seems more in line with the Aristotelian tradition which saw final and efficient causes as aspects of an explanation for a single explanandum. Second, the picture is now intuitively clearer: when a monad perceives an end, this perception serves as

Furthermore, Leibniz's arguments for substantial forms suggest that his going beyond body to the level of substantial forms and monads involves attributing efficient causality to that level. When Leibniz introduces substantial forms as force in the draft of the *New System*, he indicates that he is speaking of efficient causality: 'thus I find that the efficient cause of physical actions derives from metaphysics'.[61] And Leibniz's criticism of occasionalism in *De ipse natura* also makes it clear that the notion of force involves efficient causality. He objects as follows to the occasionalist view that the motions that now occur are the result of an eternal law decreed by God, a divine volition or command, and not at all of creaturely powers:

> Since that past command does not now exist, it cannot now bring anything about unless it left behind some subsistent effect at the time, an effect that even now endures and is at work . . . And indeed, it contradicts the notion of the pure and absolute divine power and will to suppose that God wills and yet produces or changes nothing through willing, to suppose that he always acts but never accomplishes anything and leaves behind no work or accomplishment at all.[62]

In Leibniz's own view, God's volition 'left some trace of itself impressed on things' and that means that 'we must admit that a certain efficacy has been placed in things, a form or force, something like what we usually call by the name of "nature", something from which the series of phenomena follow in accordance with the prescript of the first command'. But this force is what Leibniz thinks is the cleaned-up version of a substantial form. Now, if Leibniz thinks that this force only acts through final causality, the argument would be subject to an odd twist, where, without warning, he moves from the occasionalist denial of efficient causality to an affirmation of final causality. The most natural way to interpret this argument is that it is about efficient causality throughout, and so the conclusion is that forces or forms produce their effects through efficient causality.

But now the following question arises: if Leibniz allows for efficient causes at the level of monads and gives a role to final causes in the realm of

an efficient cause to produce an effect, the perception of the means. So when I think of going skating, this perception efficiently causes the perception of the act of getting my skates out.

In correspondence Lawrence Carlin has argued that my picture here is incorrect because it neglects the role of appetites. I have not had the chance to explore how the picture should be altered in light of this suggestion, which I do take seriously. But I do not think this issue affects the main line of my argument in this paper.

[61] G, iv. 472/WF, 22. [62] G, iv. 507/AG, 158.

bodies by way of God's purposes, why does he so often separate the two realms? The exclusion of final causes from the bodily realm makes sense in that actual explanations must be formulated in terms of mechanical, efficient causes; the role of final causes is indirect. But what about the repeated exclusion of efficient causes from the realm of bodies? Robert Adams has suggested that Leibniz meant to deny *mechanical* causation at the level of monads, but not efficient causality more generally.[63] And, indeed, this seems implicit in one of Leibniz's restatements of the pre-established harmony: 'I have shown that everything in body takes place through shape and motion, everything in souls through perception and appetite; that in the latter there is a kingdom of final causes, in the former a kingdom of efficient causes . . . '.[64] But perhaps this is not all there is to the story, unless we take Leibniz to overstate his point. Donald Rutherford writes that *explanations* at the level of bodies run in terms of efficient causes, at the level of monads in terms of final causes. Sometimes Leibniz states the point in terms of laws, and this gives us another clue: the regularities that apply in the realm of bodies fall by their nature in the realm of efficient causes. Mechanistic laws describe how mechanical events produce other mechanical events. But in the realm of monads the laws run in terms of final causes: a monad proceeds from perception to perception by way of laws about 'appetitions, ends and means', as he puts it in the *Monadology*.

Leibniz's model is voluntary action, where an intelligent agent acts on a desire for a certain result and perceives her ends as good. Leibnizian monads do not generally engage in full-blown voluntary action. Much of what happens even in an intelligent monad does not reach that level; for Leibniz only some of my perceptions are conscious and intelligent. Nevertheless he wants to apply the model of final causation across the board. How should we understand this?[65] I would suggest the following. For voluntary action, full-blown final causality applies to the monads themselves in virtue of *their* knowledge of ends. But elsewhere only 'nature teleology' applies where God's knowledge of ends is part of the account.[66] 'Natural teleology' is like the activity of Aristotelian natural agents. The monad strives for

[63] Robert Adams, 'Moral Necessity', in Jan Cover and Donald Rutherford, *Leibniz: Nature and Freedom* (New York: Oxford University Press, 2005), 186. [64] G, vii. 344/AG, 319.

[65] For extensive discussion of these ideas, see Simmons, 'Changing the Cartesian Mind'.

[66] For this term and discussion, see Donald Rutherford, 'Leibniz on Spontaneity', Cover and Rutherford, *Leibniz: Natural and Freedom*, 156–80.

ends and has immanent teleology. In line with Descartes's analysis of immanent teleology, according to Leibniz, it has perception, cognition of the ends—but not *as* ends or good.

Conclusion

Leibniz is remarkable among early moderns for the important place he gave to final causation in his system. He stands in contrast with Descartes when he agreed with the Aristotelian scholastics and other early moderns who regarded God's purposes as relevant for understanding nature. But Leibniz went further than other early moderns: like the Aristotelians, he saw immanent teleology as fundamental to understanding the true nature of genuine causal activity, and he accepted Descartes's claim that immanent teleology requires cognition.[67]

[67] This paper has benefited considerably from helpful comments from Robert Adams, Sukjae Lee, Paul Lodge, and especially Lawrence Carlin. It is a real pleasure to contribute to this volume in honor of Robert Adams, to whom I owe a great debt for his marvelous role in my life. Bob was a terrific dissertation advisor, and has ever since been a great friend, source of support, and inspiration.

8

Does Efficient Causation Presuppose Final Causation? Aquinas vs. Early Modern Mechanism

PAUL HOFFMAN

Aquinas embraced the bold claim that there can be no efficient causation without final causation—therefore, there can be no movement from one place to another without a final cause. Early modern philosophers responded with a tough counter-example: inertial motion, which apparently has no final cause. This helps to explain how early modern philosophy defeated Thomism, the last gasp of Aristotelian teleology in physics—or at least many people think that it does. But this familiar tale of triumph assumes a specific reading of Aquinas's understanding of final causation as well as a specific reading of the early modern alternative.[1] I shall argue against both, and in so doing will try to make the case that Aquinas's argument succeeds, given a stripped-down understanding of final causation that is sufficiently robust to be of philosophical significance and that should prompt us to reconsider our current understanding of what counts as a teleological explanation.

Aquinas offers a short argument setting out his view in *Summa Theologica*, IaIIae, Q1, a2.

Matter does not attain form except insofar as it is moved by an agent, for nothing brings itself from potency to act. But an agent does not move except from intention of an end; for if an agent were not determined to some effect it would not do this

[1] This introduction is almost entirely the work of Bonnie Kent.

rather than that. Therefore, to produce a determinate effect it must be determined to something certain which has the nature of an end.

I take Aquinas to be arguing as follows. In order to do anything an agent has to do something in particular. But an agent can do something in particular only if it is determined to one particular thing as opposed to some other particular thing. And to be determined to a particular thing is to have that thing as an end. So the idea here is that efficient causation requires that the cause be determined to a particular effect, which in turn entails that the cause has that effect as an end.

It is important to be clear that when Aquinas describes a cause as being determined to a particular effect he is not implying that it is determined that the effect will occur. So he explains elsewhere that

> in inanimate beings, the contingency of causes arises from imperfection and deficiency: because by their nature they are determined to one effect, which they always produce, unless there be an impediment due either to weakness of power, or some extrinsic agency, or indisposition of matter. For this reason natural causes are not indifferent to one or other result, but more often produce their effect in the same way, and seldom fail.[2]

So accidents can happen. An agent can intend one end and yet some other effect results. But the point is that in order to do anything at all, even something unintended, the agent has to be intending something in particular.

It is also important to understand what Aquinas means when, in the original quotation, he links the notion of a cause being determined to a given effect with its intending that effect. By intending an effect, Aquinas means tending to that effect, as is made clear in another passage in which he asserts that both the action of the agent bringing about the change (the mover) and the movement of the patient undergoing the change (the movable) *tend* to the end:

> Intention, as the name indicates, signifies tending to something. Now both the action of a mover and the movement of the movable tend to something. But it is due to the action of a mover that the movement of the movable tends to something. Hence intention primarily and principally belongs to the one that is the mover to an end . . .[3]

[2] *Summa Contra Gentiles* (SCG), 3a, ch. 73.

[3] *Summa Theologica*, IaIIae, Q12, a1.

Furthermore, he asserts even more clearly in *De Principiis Naturae*, chapter 3 that: 'Therefore it is possible for a natural agent to intend without deliberating about it. To intend in this way is nothing more than to have a natural inclination toward something.' So Aquinas's view, as I read him, is that the fact that something tends to some particular effect rather than another is sufficient for that effect to count as an end.

There is no explicit requirement in Aquinas's argument that the end be an endpoint or terminus. There is no requirement that the end be something good. There is no requirement that the agent act for the sake of the end or in order to achieve the end. There is no suggestion that the agent's doing something now is explained by the fact that it will lead to some future outcome. Instead, Aquinas is arguing that if cause C is determined to a particular effect E as opposed to some other particular effect, then that by itself is sufficient for E to have the nature of an end. Thus I would infer from this argument that at its core the notion of final causation for Aquinas does not depend on the assumption that motion or change presupposes an endpoint or terminus, nor does it depend on the end being good, nor need there be a purpose. Aquinas's point as I read him is that in virtue of being determined to a particular effect an efficient cause is aimed at that effect rather than other effects.

One might well object first that it is built into the very notion of final causation that there be an end that is a good or at least whose achievement counts as a purpose; and, second, that surely Aquinas links the notion of final causation not only to an end that is a good but also to its being constitutive of an end that there be an intelligence that intends it. In response to these objections it is crucial to distinguish in Aquinas between his full-bodied notion of final causation and his core notion. This distinction is implicit in the argument from the *Summa Theologica* on which I am focusing, but it is revealed much more clearly in chapters 2 and 3 of Book IIIa of the *Summa Contra Gentiles*. In the second chapter he argues that every agent acts for an end. In the third chapter he argues that every agent acts for a good. Thus he clearly thinks that the notion of acting for an end is logically prior to that end's being a good.[4]

In addition to the evidence from Aquinas's argumentation in the *Summa Contra Gentiles*, my claim that Aquinas's core notion of final causation is

[4] I am indebted to Bonnie Kent for this point.

this stripped-down version is also partly based on the assumption that the most fundamental feature of final causation for Aquinas is that it is the most fundamental of the four causes. And it would seem that in order to establish that efficient causation presupposes final causation he needs to have a stripped-down version of final causation. Were it essential to final causation that the end be an endpoint, a goal, or a good, then it is hard to see that a plausible argument could be generated to show that efficient causation presupposes final causation.

To provide some perspective on my strategy for understanding Aquinas's account of the relation between final and efficient causation, it is worthwhile to consider a rival strategy. John Carriero agrees with me that the most fundamental feature of final causation for Aquinas is that it is the most fundamental of the four causes.[5] However, his response is the exact opposite of mine. While I am proposing that in order to make sense of this view we should interpret Aquinas as relying on a stripped-down notion of final causation that I am calling his core notion of final causation, Carriero proposes that we interpret Aquinas as having a souped-up version of efficient causation.[6] Even though our strategies are opposed, I'm sympathetic to his approach and find what he says highly illuminating. It seems to me that when confronted with an argument for such a basic and yet foreign-sounding claim, such as the claim in question that efficient causation presupposes final causation, one can learn by pushing it in different directions. I hope that the way I am pushing it will also turn out to be productive.

In Carriero's view, Aquinas's notion of efficient causation is souped up because Aquinas thinks all causation requires first, that there be a movement from potentiality to actuality; second, that the effect the actuality produced be a terminus or endpoint;[7] and third, that the endpoint be a good.[8] In making such an argument Carriero is clearly appealing to Aquinas's full-bodied notion of final causation. Now I have argued that Aquinas's core notion of final causation does not require that the end be a good. However, I would argue further that the very passage from Book 3a, ch. 2 of the *Summa Contra Gentiles* that Carriero cites in support of the first two constraints shows that even Aquinas's full-bodied conception of

[5] John Carriero, 'Spinoza on Finality Causality', in Daniel Garber and Steven Nadler (eds.), *Oxford Studies in Early Modern Philosophy* (New York: Oxford University Press, 2005), ii. 113.
[6] Ibid., 106, 121. [7] Ibid., 109. [8] Ibid., 115.

final causation allows exceptions to the second. That is, Aquinas does not require that the end be an endpoint.

> For in those things which clearly act for an end, we declare the end to be that towards which the movement of the agent tends: for when this is reached, the end is said to be reached, and to fail in this is to fail in the end intended; as may be seen in the physician who aims at health, and in a man who runs towards an appointed goal. Nor does it matter, as to this, whether that which tends to an end be cognitive or not: for just as the target is the end of the archer, so is it the end of the arrow's flight. Now the movement of every agent tends to something determinate: since it is not from any power that any action proceeds, but heating proceeds from heat, and cooling from cold; wherefore actions are differentiated by their active principles. Action sometimes terminates in something made, for instance building terminates in a house, healing ends in health: while sometimes it does not so terminate, for instance, understanding and sensation. And if action terminates in something made, the movement of the agent tends by that action towards that thing made: while if it does not terminate in something made, the movement of the agent tends to the action itself. It follows therefore that every agent intends an end while acting, which end is sometimes the action itself, sometimes a thing made by the action.

The cases that do not meet Carriero's second constraint are cases such as understanding and sensation where the end is the action itself. Aquinas contrasts these cases with those in which the action terminates in something made. Elsewhere Aquinas draws a similar distinction between the sun's shining (*lucere*), which he says is an operation that remains in the sun, and the sun's illuminating (*illuminare*), which he says is an action that goes out to an exterior thing and changes it.[9] What he has in mind is Aristotle's distinction between activities on the one hand and changes or motions on the other. It is only changes or motions that have an endpoint, but it doesn't follow that an activity is not itself an end, even if it has no endpoint.

In light of the evidence first, that Aquinas does make use of a stripped-down notion of final causation in arguing that efficient causation presupposes final causation, and second, that even his full-bodied notion of final causation does not require the second constraint attributed to it by Carriero, I think we can safely pursue my approach without fear

[9] Aquinas, *Disputed Questions on Truth*, i, Q8, a6.

that is it entirely lacking plausibility. It is thus Aquinas's core notion of acting for an end that is logically prior to that end being a good that I want to consider in relation to early modern mechanism.

If Aquinas's argument, as I have construed it, works, then it would seem to be relatively easy to show that all locomotion involves final causation. First, something cannot move unless it moves in a determinate direction. Moreover, something cannot move in a determinate direction without being determined to that direction.[10] But to be determined to a determinate direction is to have that direction as an end. Thus all locomotion is teleological. Second, following Descartes and Newton, we still endorse the view that a body moving in a given direction has a tendency to continue to move in that direction. But, if Aquinas is right, to tend to move in a given direction is to have motion in that direction as an end. So there are really two considerations at work here. The first is that the linking of the action of an efficient cause with a particular effect as opposed to other particular effects entails that the effect is an end. The cause, in virtue of being determined to a given effect, is aimed at the effect. The second is that some effects themselves involve tendencies. Inertial motion involves a tendency not in the sense that a body tends toward an endpoint or terminus, but that it tends to one direction rather than another.

What shall we make of this argument? Inertial motion is not now, nor was it by its initial proponents, considered to be teleological. Indeed, one might well be inclined to think that the introduction of the notion of inertial motion in the seventeenth century was the key element in the demise of Aristotelian teleologically based science. One might defend this view by arguing that inertial motion involves a mere tendency to move in a given direction, and a tendency is not by itself sufficient for the existence of a final cause. Instead, the existence of a final cause requires more than a tendency to a certain outcome, it requires, at the very least, striving for a certain outcome. We might put this point making use of the notion of aiming. I have attributed to Aquinas the view that a cause is aimed at a given effect so long as it is determined to that effect, that is, that it tends to

[10] To avoid complications I am assuming what Aquinas would consider the normal case in which there is no impediment. To cover the complete range of cases this premise would have to be revised to read 'something cannot move in a determinate direction without being determined to some direction or other'.

that effect, but one might reply that a cause is not aimed or is not aiming at a given effect unless it is striving towards it.

Let me cloud the waters, if they aren't cloudy already. If we look at the views of Descartes, Newton, and Spinoza, it is very hard to ascertain first, what sort of distinction they drew, if any, between the notions of tending toward and striving toward, and second, if they did draw such a distinction, whether they would have denied that bodies strive to maintain rectilinear motion. The primary source of the mystery is the phrase *quantum in se est*. Descartes makes use of it in formulating his law of inertia; he is followed in this by Newton, and Spinoza uses it in his doctrine of universal conatus.

> Each thing, in so far as it is simple and undivided, always remains in the same state, *quantum in se est*, and never changes except as a result of external causes.[11]

> The *vis insita*, or innate force of matter, is a power of resisting, by which every body, *quantum in se est*, continues in its present state, whether it be of rest, or of moving uniformly forwards in a straight line.[12]

> Each thing, *quantum in se est*, strives to persevere in its being.[13]

The phrase is obviously intended to do some important work. But what? It gets translated in various ways, which makes it all the more difficult to understand what work it is supposed to be doing. The most literal translation is 'in so far as it is in itself'. One common translation is 'in so far as it can by its own power'. I. Bernard Cohen has made a powerful case, however, that it is best translated as 'according to its nature'.[14] All three of these translations suggest that there is something internal to the thing in virtue of which it remains in its same state or, in the case of Spinoza, perseveres in its being.

It is important to emphasize here that Descartes is widely misread as supposing that there is nothing internal to a body in virtue of which it continues in the same state, and as thinking that there is no sense in which bodies are active. Instead, these tendencies can be explained entirely by reference to God's will without attributing force or anything else internal to bodies. Explicit evidence that Descartes thinks we do need to attribute a

[11] *Principles*, ii. 37; AT, viiia, 62; CSM, i. 240–1.
[12] *Principia*, C i. 2, Definition III. [13] *Ethics*, iii. P6.
[14] I. Bernard Cohen, '"Quantum in se est": Newton's Concept of Inertia in Relation to Descartes and Lucretius', *Notes and Records of the Royal Society*, 19 (1964), 131–55.

force internal to bodies to account for their behavior is found in a neglected passage from his correspondence. In a letter to Mersenne, dated 28 October 1640, he states:

> He [Father J. Lacombe] is right in saying that it was a big mistake to accept the principle that no body moves of itself. For it is certain that a body, once it has begun to move, has in itself for that reason alone the force to continue to move, just as, once it is stationary in a certain place, it has for that reason alone the force to continue to remain there. But as for the principle of movement which he imagines to be different in each body, this is altogether imaginary.[15]

This passage makes it clear that Descartes's objection to Aristotelian internal principles of movement in bodies is not that he thinks bodies have no internal tendencies, but rather that in failing to recognize that all matter has the same nature the Aristotelians have failed to recognize that all matter in motion has the same one tendency—namely, to continue moving in a straight line.[16]

Is the phrase '*quantum in se est*' doing more work than indicating that there is an internal source of a thing's preserving its state? The 'in so far as it can' translation, at least to my ear, suggests that the thing is striving. So if that is the best translation, then it would imply that both Descartes and Newton thought that inertia does involve a body striving to maintain its present state. But, if instead, 'according to its nature' is the best translation, so that what is being claimed is that a body remains in the same state because of its nature, it seems less clear that striving is involved, although, on the other hand, I do not think that striving is being excluded. If by 'nature' what Descartes and Newton have in mind is an innate or natural force, so that it is in virtue of an innate or natural force that a body preserves its state, then it is tempting to think that in acting a body is making an

[15] AT, iii. 213; CSMK, 155.

[16] It is important to distinguish between natural tendencies and internal tendencies. According to Descartes the only natural tendency bodies have is the general tendency to remain in their same state. This is why a body at rest tends to stay at rest and a body in motion tends to continue moving. Descartes thinks that a body in motion tends to continue moving in a straight line because all motion is rectilinear. But, since neither being at rest nor being in motion is natural to a body, the determinate tendency of a resting body to stay at rest and that of a moving body to continue moving, while internal to the body, are nevertheless not natural. This has the important consequence, pointed out by Jeffrey McDonough, that according to Descartes's system we have to observe a body's current state in order to ascertain its determinate tendency. This contrasts with the Aristotelian system according to which knowing a body's nature is sufficient to ascertain its determinate tendency.

effort, that is, striving. However, this cannot be what Newton has in mind, because he explicitly identifies the force of inertia with the 'inactivity of the mass'.

Michael Della Rocca has made a convincing argument that the phrase '*quantum in se est*' as used by Descartes is meant to flag the notion of striving. He notes various passages in which Descartes uses the Latin equivalent of 'strive', and its variants, as a substitute for the phrase '*quantum in se est*'. So, for example, Della Rocca quotes the following passage:

> When I say that the globules of the second element strive (*conari)* to move away from the centres around which they revolve, it should not be thought that I am implying that they have some thought from which this striving (*conatus*) proceeds. I mean merely that they are positioned and pushed into motion in such a way that they will in fact travel in that direction, unless they are prevented by some other cause.[17]

Della Rocca goes on to argue that Descartes is offering a deflationary account of striving, that is, he argues that the notions of striving and tending are equivalent for Descartes.[18] I think this is a mistake. It seems to me that striving for Descartes is more than mere tending, it is tending in which there is an internal source of the tending. That is, only if something tends *quantum in se est*, is it striving. To elaborate, I would argue that it is not sufficient for the source of body B's tending to x to be internal that its current state s is such that it will do x unless prevented. Admittedly in that case B's doing x is a function of its being in internal state s, but I think the crucial question is, what accounts for that function's holding? If it is the case that it is entirely due to God's action that B's being in s results in its doing x, then I would say the source of the tendency is external. As far as I know, Descartes never mentions this scenario, but I see it as lurking in the background. In my view the source of the tendency is internal only if something internal to the thing contributes to the explanation of why its doing x follows from its being in internal state s. That is, I want there to be something internal to the thing that looks like a force or efficient cause. So it is in this latter scenario that I think there is striving

[17] *Principles of Philosophy*, iii. 456; AT, viiia. 108; CSM, i. 259. See Michael Della Rocca, 'Spinoza's Metaphysical Psychology', Don Garrett (ed.), *The Cambridge Companion to Spinoza* (Cambridge: Cambridge University Press, 1996), 195.

[18] Della Rocca, 'Spinoza's Metaphysical Psychology', 195–6.

on the part of the body, but not in the former scenario where God is doing all the work of taking the thing from the internal state s to the doing of x.

One might object at this point that striving involves conscious effort, or, at least, one might object that Descartes thought striving involves conscious effort, so that in explaining what he means by striving—that bodies 'are positioned and pushed into motion in such a way that they will in fact travel in that direction, unless they are prevented by some other cause'—Descartes is not offering a deflationary account of striving, but rather an eliminative account. And it might be claimed that textual evidence to support this eliminative interpretation is found in his argument for rejecting the Aristotelian account of gravity:

> But what makes it especially clear that my idea of gravity was taken largely from the idea I had of the mind is the fact that I thought that gravity carried bodies towards the centre of the earth as if it had some knowledge of the centre within itself. For this surely could not happen without knowledge, and there can be no knowledge except in a mind.[19]

Certainly Descartes is asserting here that an internal force cannot carry a body to an endpoint without knowledge of that endpoint. But is he also asserting that to suppose there is a force in a body carrying it in a straight line also implies that the body has knowledge? I do not think he is making this further assertion. So I read Descartes as maintaining that as long as there is a force in a body in virtue of which it tends in a certain direction, that is sufficient to say it is striving. That is, I think his considered opinion is that striving need not arise from thought.[20]

This argument that inertial motion as conceived at least by some early modern mechanist philosophers involves striving could be taken to have completely opposite implications. On the one hand, it might be taken to

[19] AT, vii. 442; CSM, ii. 298.

[20] So in the passage discussed above in which Descartes asserts, 'When I say that the globules of the second element strive (*conari)* to move away from the centres around which they revolve, it should not be thought that I am implying that they have some thought from which this striving (*conatus*) proceeds. I mean merely that they are positioned and pushed into motion in such a way that they will in fact travel in that direction, unless they are prevented by some other cause'. I take him to be arguing only that striving does not require thought. I do not read him as making the further argument that the globules can still be said to strive even if the explanation of why they will in fact continue to travel away from the centers if not impeded need not invoke a force internal to the globules.

show even striving for an end is not sufficient for final causation. I will return to this issue later. On the other hand, to someone committed to the view that striving for an end is sufficient for final causation, it might be taken to show that Descartes is wrong to conclude that inertial motion does not involve final causation. I am inclined to think that Aquinas would say that Descartes was wrong to conclude that inertial motion as he conceived it does not involve a final cause, at least understood according to the stripped-down core notion. To elaborate, it does seem plausible to argue that in conceiving inertial motion as something that does not have an endpoint, the early modern mechanist philosophers conceive of it as what the Aristotelians would have considered to be an activity rather than what they would have considered to be motion. That is, they have assimilated inertial motion to activities such as sensing and understanding and shining insofar as none of them has an endpoint. But, as noted above, it does not follow from that transformation that inertial motion lacks the character of being an end.

Still another response would be to argue that in spite of Descartes's references to forces in bodies and Newton's use of the term 'force of inertia' (*vis inertiae)*, which is considered by many to be a contradiction in terms, and in spite of their common use of the expression *quantum in se est*, the real significance of the introduction of the concept of inertia is that the continuation of motion does not require a force at all. Only changes in motion require forces. That is, it is not that the continuation of rectilinear motion is being reconceived so that it is viewed as an activity rather than as a change, but rather it is being reconceived as a state, on a par with a thing's shape. And just as no efficient cause is required for something to maintain its shape (in the absence of external forces), no efficient cause is required for something to maintain its state of rectilinear motion in the absence of external forces. Furthermore, to say that inertial motion does not require a force is tantamount to saying that it does not require an efficient cause. So the case of inertial motion is not really relevant to the question of whether efficient causes presuppose final causes.

In light of these various responses, especially the last one that inertial motion as conceived by Descartes and Newton has no efficient cause, I suspect that many readers will find it inconclusive whether inertial motion constitutes a counter-example to Aquinas's claim. Perhaps it would be better to focus on cases of change of motion, which are cases in which force

does come into play and which are still considered by most contemporary philosophers, I presume, not to involve final causes.

Newton's other main concept of force, besides the force of inertia that is innate to the moving body, is the force proportional to a change in motion. He refers to this force as impressed force. It would appear to be external to the moved body in its origin, since it is defined as an action exerted on a body. I have to confess that I do not have a very good understanding of this notion of force, but I don't think we think of this force, this action, as having any connection with the notion of striving. However, it is interesting that Newton himself defines the force as being 'exerted upon a body, *in order to* change its state' [*Vis impressa est actio in corpus exercita, ad mutandum eius statum vel quiescendi vel movendi uniformiter in directum*] which certainly seems to have a teleological ring to it. It is also noteworthy that Newton asserts that 'this force consists in the action alone and does not remain in the body after the action'. Since Newton identifies the action with the impressed force, and impressed force was understood to be a force received in the moved body, the strong suggestion of these remarks is that Newton views the action that brings about motion as being located in the moved body rather than in the agent causing the motion. So Newton's account of impressed force has two Aristotelian echoes. The first is that the action has a purpose and the second is that the agent's action is located in the patient, that is, the thing acted on. Still, even if Newton was thinking of impressed forces along Aristotelian lines, we do not think of forces in that way. We do not think of them as actions located in the moved body nor do we think of them as exerted in order to do something.

Nevertheless, I think we do think of impressed forces as having determinate effects that we can specify. So this brings us back to Aquinas's original argument. Is the mere fact that we can specify these effects sufficient to make them ends and thus to justify Aquinas's view that efficient causation presupposes final causation?

Perhaps what is at stake is the underlying explanation of why something is determined to one effect rather than another. That is, what is at stake is what underlies the laws of nature. There would seem to be three possibilities. One is that adopted by Spinoza, that effects follow from causes by a kind of conceptual necessity in the same way that properties of a triangle follow from its essence. On such a picture it does not seem unreasonable to deny

that the cause is aimed at the effect. However, it is not clear to me that on such a picture we have to deny that the cause is aimed at the effect. If the cause is, as it were, locked into the effect as a matter of conceptual necessity, why not say it is aimed at the effect? It might seem odd, however, on such a picture, to characterize the cause as striving toward the effect. If it is a matter of conceptual necessity that cause C is directed to effect E, that it will tend to have E as a result, that is, in the absence of other countervailing causes E will occur, then there does not seem any need for striving. But this points at most to an inconsistency in Spinoza, for whom striving plays a central role,[21] and not to a real philosophical problem for the view that a cause can be said to be aimed at an effect to which it is connected by way of conceptual necessity. In any case, it is a very uncommon view, unfashionable since Hume, to think that there could be a conceptual connection between cause and effect.

Second, one might think, as did Leibniz, that there is an author of nature who has forged the connections between cause and effect and that there is a sufficient reason for the connections being made as they are. This of course would be a teleological conception, and, in Leibniz's case, since the reason has to do with its being the best of possibilities, it is teleological in the fullest sense.

Third, one might think that it is a matter of brute fact, not subject to further explanation, that the fundamental laws of nature are as they are. And one might try to argue further that if it is merely a brute fact that any cause C is connected to an effect E, then there are no final causes. However, I'm tempted to think that at least part of the point of Aquinas's argument is to undercut this last move. Even if it is nothing but a brute fact that cause C is connected to E, the mere fact that C is connected to E rather than to E′ entails that C is aimed at E. So the idea is that being aimed at an effect is sufficient for final causation; it is not necessary that the cause strive towards that effect.

Now suppose we grant that Aquinas is right and we can admit that every cause is aimed at an effect and in that sense the effect is an end, and so in that minimal sense of final causation, efficient causation presupposes final causation. Does such a stripped-down notion of final causation have any

[21] Michael Della Rocca has pointed out that we need not suppose that Spinoza is being inconsistent. We could argue in his defense that the fact that things strive to persevere in being is indeed part of the explanation of why effects follow by conceptual necessity.

philosophical interest? I think it does. I think it is a significant insight on Aquinas's part that specification is equivalent to aiming, in other words, to specify a cause's effect is tantamount to saying that the cause is aimed at the effect, and it really does not matter whether there is an explanation underlying the specification or whether it is simply a brute fact. I am suggesting, in other words, that one lesson of Aquinas's argument is that we are mistaken in thinking that there is nothing teleological about laws of nature that say that cause C will result in effect E, other things being equal. The sort of tendency expressed by such laws has the implication that C is aimed at E, and that is sufficient to give the laws a teleological character.

This, of course, is an extremely controversial claim. Many contemporary philosophers would in contrast cite the behavior of a heat-seeking missile as behavior that is genuinely teleological, on the ground that teleology requires that something reorient itself to its target if it is knocked off course. On the weak view I am deriving from Aquinas, a thing's behavior has teleological character provided only that it would act in a certain way provided nothing interferes (or, alternatively, a causal sequence counts as teleological so long as the cause would have a given effect provided nothing interferes). I do not have a further argument to confront skeptics who doubt that this sort of conditional is sufficient to ground the claim that the thing is aimed at that sort of behavior (or that the cause is aimed at the effect). But I am inclined to think that what is needed here is not further argument, but rather, as happened to me, a gestalt shift.

Aquinas commits himself to the stronger view that a cause is determined to, that is, is aimed at a given effect only if it almost always achieves that effect. But that condition seems unnecessarily strong. Rather, what seems correct is that a cause is determined to or aimed at an effect provided that the effect would come about 'unless there be an impediment due either to weakness of power, or some extrinsic agency, or indisposition of matter'. And it seems that it could happen that a cause is determined to a given effect even if there is an extrinsic agent that interferes most or perhaps even all of the time, because the cause would have achieved the effect had there been no such impediment.

What would count as a counter-example to Aquinas's claim that efficient causation presupposes final causation on this stripped-down core notion of

final causation? It would have to be a cause such that it is not true that the specified effect would come about unless there were an impediment. In other words, the specified effect might not come about even if there is no explanation for its not coming about. What this shows, I think, is that if we grant to Aquinas the principle of sufficient reason construed in the non-teleological sense that we must be able to give an efficient causal explanation for everything that happens, then he is home free.

One question that has come up is whether indeterministic motions such as one might find in quantum mechanics or completely random motions would count as counter-examples to the view that all motion requires that something is being aimed at. I am not sure what to say about these cases. In regard to the case of indeterministic motions, it is not entirely clear to me that a cause has to be aimed at a single effect in order to count as being aimed. So suppose that a cause (independently of any other causal influences) has various incompatible effects distributed over a probability space. I don't see why we should not say that it is aimed at all these various effects.[22] Aquinas himself would not be satisfied with this response. His view is that if a cause is going to act, it cannot be indifferent to two or more effects, but a cause whose probability space included two incompatible effects each with 50 per cent probability would seem to be indifferent between them.[23] What this reveals, perhaps, is that if we are willing to say that an indeterministic cause is aimed at its various effects, then that notion of being aimed is weaker than Aquinas's notion of a cause being determined to a particular effect because Aquinas's notion of determination precludes indifference. In regard to the second case of completely random motion, one might still try to argue that aiming is not even lost here. It seems correct both that something cannot move without heading in some direction or other and that to be headed in a direction is to be aimed.[24]

One objection that has been made is that it is impossible to sustain the distinction I have tried to draw between those effects that are aimed at and those effects that are merely the accidental result of the interaction of various causes individually tending to or aimed at other effects. So, for example, if body A is tending in a certain direction and body B is tending

[22] I am indebted here to Robert M. Adams.
[23] See SCG, 3a, ch 2.
[24] I am indebted here to Bonnie Kent.

in another direction, I have wanted to say that A's collision with B might be accidental and not something either is aiming at. But one might argue that from the broader perspective of the system containing A and B, since they would collide provided no other causes intervene, their collision is something that is aimed at.[25]

I am prepared to grant that what is accidental from the point of view of one agent might not be accidental from a more encompassing perspective involving many agents, and that there might be some comprehensive perspective from which nothing is accidental. A similar issue was of concern to scholastic Aristotelians. As Dennis Des Chene has noted, neither the principal cause (the father) nor the impeding cause was thought to intend the production of a monster, but Nature understood as the total cause was thought to intend it because it inclines towards it.[26]

None of this discussion is to deny that there are other richer conceptions of final causation and that other discussions of final causation tend to focus on these richer accounts. I've been arguing that in a minimal but still perfectly good sense final causation is present whenever the cause is aimed at the effect. In a stronger sense of final causation, the cause is not only aimed at the effect, but it strives for the effect. In arguing that Spinoza advocates final causation in still another stronger or richer sense, Don Garrett has asserted, in the tradition of my colleague Larry Wright,[27] that an explanation of a thing's behavior B is teleological if B's origin or etiology is explained by its having as a usual or expected outcome O. Garrett claims that Spinoza is committed to such explanations, for example, that humans develop with sharp teeth in front so that they can tear food. In such a definition an explanation would count as teleological even if the notion of striving plays no role in the explanans. Garrett argues conversely that the fact that Spinoza thinks finite things strive to preserve themselves 'provides an obvious avenue for explaining the behavior of singular things by appeal to the self-preserving tendency of that behavior'.[28]

[25] This is my understanding of the objection pressed upon me by Martin Schwab.

[26] Dennis Des Chene, *Physiologia: Natural Philosophy in Late Aristotelian and Cartesian Thought* (Ithaca, NY: Cornell University Press, 1996), 207.

[27] Larry Wright, *Teleological Explanations: An Etiological Analysis of Goals and Functions* (Berkeley, Calif.: University of California Press, 1976).

[28] Don Garrett, 'Teleology in Spinoza and Early Modern Rationalism', in Rocco J. Gennero and Charles Huenneman (eds.), *New Essays on the Rationalists* (Oxford: Oxford University Press, 1999), 313.

I agree that that avenue is open to Spinoza, but the key interpretive question is whether Spinoza goes down it. I see no evidence that he does. That is, I see no evidence that Spinoza argues that all (or perhaps even that any) self-preserving behavior has its etiology in the fact that it is self-preserving. So I am inclined to think, contrary to Garrett, that Spinoza and Descartes are in the same boat with respect to unthoughtful teleology.[29] Both think bodies do strive and both think that is not sufficient for final causation, but if Aquinas is correct they are wrong in that judgment.

Still another even stronger conception of final causation requires that the end in question either be a good or be viewed as a good. Indeed it has been argued recently that this condition is essential for all final causation.[30] And one might assert, as Carriero does, that Aquinas himself is committed to this strong conception of final causation. I would respond to this in two ways. First, I agree that according to Aquinas's full-bodied conception of final causation every agent does act for sake of some good. However, as I argued at the outset, he regards the principle that every agent acts for sake of some good as a different principle from the principle that every agent acts for an end, and he offers separate arguments for them. Second, Aquinas has a very broad conception of the good, so broad that fire begetting fire counts for him as an agent acting for a good. He is committed to the Aristotelian view that good and being are extensionally equivalent, that evil is only found in the privation of actuality. Thus Aquinas's conception of the good is so broad that a body's causing itself to continue moving or a body causing another body to change its state of motion would also count as acting for a good. And, since Descartes considers rest not to be the privation of motion but as a mode with as much reality as motion, even causing something to stop moving would satisfy Aquinas's criterion of an agent's acting for a good. It could therefore plausibly be argued that even leaving aside God's role in Descartes's system, his physics can still be counted as teleological on a conception of teleology according to which agents must act for a good.

My primary conclusion is that it is reasonable to read Aquinas as operating with a stripped-down conception of final causation when he argues that

[29] Ibid., 326, 332.

[30] Mark Bedau, 'Where's the Good in Teleology', *Philosophy and Phenomenological Research*, 52 (1992), 781–806.

efficient causation presupposes final causation. Yet this stripped-down conception is still of philosophical interest. It is not empty to assert that all efficient causes are aimed at something.[31]

[31] Versions of this paper were presented at: the University of California, Riverside agency workshop; University of California, Irvine; UCLA history workshop; the University of Toronto; and the New England Colloquium in Early Modern Philosophy held at Yale University. I have received many very helpful suggestions and objections. I would like to thank Bonnie Kent, John Carriero, Jeffrey McDonough, Hannah Ginsborg, Don Garrett, Michael Della Rocca, John Fischer, Neal Tognazzini, Samantha Matherne, Larry Wright, Ermanno Bencivenga, Martin Schwab, Nicholas Jolley, William Bristow, David Woodruff Smith, Sean Kelsey, Tyler Burge, Marleen Rozemond, Jennifer Whiting, Donald Ainslee, Elmar Kremer, Kenneth Winkler, Justin Broakes, James Kreines, Lisa Downing, Larry Jorgensen, and Robert M. Adams.

9

Herder and Kant on History: Their Enlightenment Faith

ALLEN WOOD

One of Kant's least attractive traits as a human being was a tendency to regard his students, followers, and protégés as disloyal to him if they departed from what he saw as the central tenets of the critical philosophy. Kant displayed this ugly trait especially toward two men—Johann Gottlieb Fichte and Johann Gottfried Herder—through whom Kantian philosophy was greatly enriched and its influence extended far more than it would otherwise have been. Both were great and original philosophers in their own right, but also touchy, difficult, impossible personalities, far more flawed in dealing with others than Kant ever was. Fichte was a better friend to the critical philosophy than Kant ever realized, however, often drawing conclusions from Kantian principles far more consequently than Kant himself did. And Herder's thought, along with Kant's, helped to revolutionize the study of human nature and history. In both cases, the personal conflicts have led people to imagine deeper philosophical divisions than I think are really there.

Herder was Kant's student, and most of what we know about Kant's lectures on metaphysics and ethics between 1762 and 1764 comes from lecture transcriptions in his hand. Herder came to prominence when his essay *On the Origin of Language* won a prize from the Prussian Academy in Berlin in 1771, but his most important contributions to the philosophy of history were the short essay *This Too a Philosophy of History for the Formation of Humanity* (1774) and the massive *Ideas for the Philosophy of the History of Humanity* (1784–91). Herder's contributions to anthropology and philosophy of history thus pre-date Kant's, and the mutual influence between the two thinkers is clearly greater than is commonly appreciated.

Kant's personal relations with Herder were complex. Kant's reviews of Herder's *Ideas* were superficially laudatory but insultingly condescending. Herder's last works, *Metacritique of the Critique of Pure Reason* (1799) and *Kalligone* (1800) were polemics against Kant. Yet in the late 1790s, Herder's *Letters on the Advancement of Humanity* included an eloquent tribute to Kant, which Lewis White Beck used to quote at the conclusion of the prefatory 'Life and Works' section in his translations of Kant.[1]

Many treatments of Herder see him mainly as a follower of Kant's eccentric counter-enlightenment friend Hamann, and an enemy of Kantian philosophy. Herder is characterized as a figure of the *Sturm und Drang*, a counter-enlightenment thinker, a nationalist rather than a cosmopolitan, a Francophobe and Germanophile, philosopher of feeling and intuition rather than reason, a cultural relativist and an opponent of moral universalism. But these images represent half-truths at most. Fred Beiser seems to me to get closer to the truth when he describes Herder's philosophy as resulting

[1] See *Prolegomena to Any Future Metaphysics* (Indianapolis, Ind.: Bobbs-Merrill, 1950), p. xxii; *Critique of Practical Reason* (Indianapolis, Ind.: Bobbs-Merrill, 1958), p. xxii; *Foundations of the Metaphysics of Morals* (Indanapolis, Ind.: Bobbs-Merrill, 1959), p. xxii; *Kant on History* (Indianapolis, Ind.: Bobbs-Merrill, 1963), p. xxviii. The quotation is from Herder, *Briefe zur Beförderung der Menschheit* (1793–7) in *Werke*, ed. B. Suphan (33 vols.; Berlin: Weidmann, 1877–1913), xviii. 324–5. Herder will be cited below by volume : page number in this edition. Where appropriate, Forster's translation will also be cited: Forster (ed.), *Herder: Philosophical Writings* (Cambridge: Cambridge University Press, 2002). 'Change of Taste' will be abbreviated VdG and cited in Forster's translation and according to the volume : page number of the only edition in which it has been published—namely, Johann Gottfried Herder, *Werke*, ed. Gaier (Frankfurt am Main: Deutscher Klassiker Verlag, 1985).

Ak	*Immanuel Kants Schriften*, Ausgabe der königlich preussischen Akademie der Wissenschaften (Berlin: W. de Gruyter, 1902–). Unless otherwise footnoted, writings of Immanuel Kant will be cited by volume : page number in this edition
Ca	*Cambridge Edition of the Writings of Immanuel Kant* (New York: Cambridge University Press, 1992–) This edition provides marginal Ak volume : page citations. Specific works will be cited using the following system of abbreviations (works not abbreviated below will be cited simply as Ak volume : page)
EF	*Zum ewigen Frieden: Ein philosophischer Entwurf* (1795), Ak, 8 *Toward Perpetual Peace: A Philosophical Project*, Ca Practical Philosophy
IAG	*Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht* (1784), Ak, 8 *Idea Toward a Universal History with a Cosmopolitan Aim, Ca*, Anthropology History and Education
MS	*Metaphysik der Sitten* (1797–8), Ak, 6 *Metaphysics of morals*, *Ca*, Practical Philosophy
P	*Prolegomena zu einer jeden künftigen Metaphysik* (1783), Ak 4 *Prolegomena to Any Future Metaphysics, Ca*, Theoretical Philosophy after 1781
RH	Rezensionen von Herders Ideen, Ak, 8 Reviews of Herder's Ideas, *Ca*, Anthropology, History and Education

from a struggle between the twin influences of Hamann and Kant, and also when he concludes that if either of Herder's teachers could be said to have won the struggle, then it was Kant.[2] No doubt Herder and Kant disagreed about a number of things in philosophy.[3] But on the philosophy of history, which is my present theme, I think we can properly appreciate the real differences only if we begin with the recognition of how much Kant and Herder have in common, and how even their quarrels rest on more basic points of agreement.

The main point on which Kant and Herder agree, I will argue, is that it is rational to look at human history as exhibiting a kind of natural purposiveness, like that found in organisms rather than that found in intentional human actions. Both philosophers, I will argue, should be seen as grounding their philosophy of history on what could be described as a faith in human progress. This is a faith that is characteristic of (though by no means universal in) the Enlightenment. Sometimes Enlightenment faith takes a religious form, sometimes a purely secular form, and sometimes a mixture of the two (as in both Kant and Herder).

In our time it seems to be fashionable to criticize or condescend to Enlightenment views of historical progress, as though they represented some sort of naïveté that we—who have lived to see such twentieth-century events as the Nazi Holocaust, the bitter fruits of European imperialism, and the collapse of socialism—can now recognize as a kind of illusion. Post-modernism now dismisses such Enlightenment views as 'meta-narratives,' and many religiously inspired views have come to see them as exhibiting a kind of secular humanist hubris. For this reason, after presenting the views of Herder and Kant, I will conclude with some reflections on their faith in historical progress. To this end I will call upon some of Robert Adams's

[2] Frederick C. Beiser, *Enlightenment, Revolution and Romanticism* (Cambridge, Mass.: Harvard University Press, 1992), ch. 8.

[3] Michael Forster puts forward the interesting thesis that Herder derived his main philosophical opinions from Kant—but the pre-critical Kant of the 1760s, not the critical Kant of 1781 and after, with whom he had some profound differences. As Forster sees it (*Herder: Philosophical Writings*, pp. xi–xiv), Herder accepts Kant's early sentimentalist (or, as Forster puts it, 'noncognitivist') views in ethics, his '(Pyrrhonist-influenced) skepticism about metaphysics', and a brand of empiricism that the pre-critical Kant is supposed to have held—all of which positions are starkly opposed to the Kant of the later critical philosophy. Here I would sooner question these characterizations of the early Kant than Forster's attribution of them to Herder. And they certainly do point to important disagreements between Kant and Herder on many philosophical issues. But I repeat that my present theme is only the philosophy of history, where the disagreements are sometimes thought to be greatest, but where I think they are far less than is usually supposed.

thoughts about *faith*—especially the kind of 'moral faith' he discusses in the last chapter of *Finite and Infinite Goods*. I will try to use Adams's account to make the case that the Herderian and Kantian views of history are neither objectionably naïve nor hubristic, but quite reasonable. Enlightenment faith in historical progress, I will argue, is a kind of rational faith (in Adams's sense of 'faith') that is rationally necessary for us as often as we seek a rational understanding of human history and of our own actions and strivings as part of it.[4]

Herder's Historical Manifesto

Kant is well known for the range of his philosophical interests and accomplishments. His philosophy revolutionized virtually every field in philosophy: metaphysics, epistemology, philosophy of mind, philosophy of science, moral and political philosophy, aesthetics, philosophy of religion, and (as I am arguing here) philosophy of history. It is much less well known that very much the same could be said of Herder in many fields of the human studies, such as philosophy of mind, philosophy of language, philosophy of action, the methodology of history and anthropology, and both biblical and literary interpretation. The theoretical revolutions of Schleiermacher in hermeneutics and Wilhelm von Humboldt in linguistics were built directly on Herder's achievements.[5]

In 1774, Herder published *Auch eine Philosophie der Geschichte zur Bildung der Menschheit*—a emotional and declamatory little essay, wildly ambitious in scope, rich and original in content, by turns impassioned and sarcastic in tone. The standard reading takes it to be an anti-Enlightenment polemic. Herder's essay is filled with indictments of his own age, regarded as an age of Enlightenment, an age of philosophy. It is an age of abstract intellect, in which the head is divided against the heart, in which a superficial, artificial, mechanical way of thinking has cut people off from their deeper humanity. It is an age of skepticism, of religious unbelief, of abstract cosmopolitanism,

[4] I owe to comments by Samuel Newlands on an earlier draft of this paper the challenges that led me into this discussion, and into using Adams's conception of faith for this purpose.

[5] For a good account of the range of Herder's accomplishments and influence, see Forster, 'Johann Gottfried Herder', *The Stanford Encyclopedia of Philosophy* (Summer 2007 edn.), ed. Edward Zalta <http://plato.stanford.edu/entries/herder/> cited hereafter as 'Forster, *SEP*'.

in which the values of patriotism and cultural rootedness have been lost. Finally, this is an age of narrow-minded complacency, that looks down from the supposed height of its philosophical wisdom on all earlier ages, self-conceitedly regarding itself as the final goal of human history. All these traits are symptomatic, Herder seems to be saying, of an exhaustion of the powers of life.

By contrast, Herder praises the virtues of past times, especially those to which he thinks his age condemns and to which it feels superior—the patriarchal religion of the Orient, the priestcraft of Egypt, the patriotism of Greece and Rome, the dominance of religion over all spheres of life in the long era of Christianity that preceded the Renaissance and Reformation. Much of this polemic against his own time is continuous with the works of literary criticism that preceded his historical manifesto. In them Herder argued that ancient languages, such as that of Homer and the Bible, were more poetic than modern languages, which had become too intellectualized, too rule-governed, too cut off from everyday life. He celebrated the variety of folk-culture over the classicism that saw beauty only in the imitation of a narrow range of privileged models taken from Greco-Roman culture.[6]

Herder's essay is often read exclusively as I have just been presenting it, partly because there are many who find the position it puts forward, so interpreted, highly attractive. According to these interpreters, Herder sees through the arid intellectualism of us philosophers, recognizes the shallow elitism that hides behind the Enlightenment's façade of egalitarianism. He is in touch with our deeper humanity, which is denied by the shallow scientific rationalism of the eighteenth century, and so on and so forth. Herder is then seen as the harbinger of the Romantic movement, the earliest perceiver of the truths articulated only recently by multiculturalism and postmodernism.

[6] Above all, he defended German culture against domination by the French—a feature of his work that later led the Nazis to claim him, along with Fichte, as one of their own. In both cases, of course, such a claim is false to the point of obscenity, but it is especially distorting in the case of Herder, whose supposed 'nationalism' involved an intense hostility to most of the things—such as state and military power—that that term connotes. Besides, the proper calling of German culture in Herder's view was always thought, poetry, and teaching, and for him 'German' culture included even Shakespeare and the supposed Celtic folk poetry of Ossian (which Herder regarded as authentic, even after cooler heads, such as Hume, had immediately recognized MacPherson's crude forgeries for what they were). For another thing, Herder's Germanophilia was always cultural, never political, and least of all military.

I would not want to deny that Herder's thought does anticipate much in these later movements or that it has exercised much influence on them (much of it indirect and even now insufficiently acknowledged). But this reading of Herder still seems to me profoundly wrong, not only in its view of Herder but even more in its orientation to modernity. To begin with, it leaves people like me far too easy a way out—I mean those of us who embrace the Enlightenment precisely *because* it was an age of skepticism and unbelief, *because* it rejected patriotism and sentimentalism, because we think the intellectual honesty of natural science exhibits more vitality than the loathsome emotionalism of preachers and poets—the latter accurately pegged by Hume as 'liars by profession'.[7] To read Herder simply as 'counter-Enlightenment' merely reduces the issues between us and him to matters of taste—bad taste on his part, good taste on ours.

Then too, on this reading we can reject a lot of what Herder seems to be saying about the Enlightenment as simply a gross mischaracterization. Which Enlightenment thinkers are supposed to be self-complacent, thinking that the whole of history leads up to them as its pinnacle and final end? Some, such as Voltaire and Mendelssohn, don't seem to believe in historical progress at all. Those who, like Kant, do cautiously entertain the idea of historical progress—more cautiously, in fact, than Herder did—focus their hopes on the future, not on the historical present—exactly as Herder himself does in the closing pages of *This Too a Philosophy of History*.

As this last point makes clear, the exclusively counter-Enlightenment reading also makes it all too easy to cite Herder against himself, since he himself is a partisan of Enlightenment values just about everywhere it counts. Herder's religious faith turns out to be that of a typical German freethinker of the time—highly personal and heterodox, laced with Spinozistic heresies and idiosyncratic, non-literal readings of the Bible. Herder is an advocate of toleration and open discussion in all matters, anti-authoritarian, anti-militaristic, anti-imperialistic, egalitarian in his social, economic, and political convictions. Herder's chief aim in the human sciences was to take better account of the empirical facts, to expand and adjust scientific method, not to reject it. What we need, therefore, is a reading of Herder that explains why he polemicizes so ardently against

[7] Hume, *Treatise of Human Nature*, ed. Selby-Bigge (Oxford: Clarendon Press, 1967), 161.

some prominent representatives of the Enlightenment, such as Voltaire and Kant, and shows us what he wants to add to the Enlightenment, while acknowledging the equally plain fact that when the chips are down, he is himself a part of it.

Herder's Greatest Contribution

The best way to reach a better reading of Herder's 1774 historical manifesto is to begin with its most revolutionary idea. This is Herder's new way of looking at cultures and ages different from our own. Every age of human history, Herder claims, must be understood from within, in terms of its own 'way of thinking' (*Denkungsart)*, which includes its own ways of experiencing and feeling, as well as of living and acting. Each age is unique, to be understood first of all simply in terms of itself and its inner organic forces, like a living thing or a work of art, and not mechanistically, as simply an instance of some general laws. As he puts it later in the *Ideas*: 'The feelings and inclinations of human beings are everywhere conformable to their organization and the circumstances in which they live; but they are everywhere swayed by custom and opinion'.[8]

In the first section of *This Too a Philosophy of History* Herder reviews the main cultures of the ancient world and attempts to characterize each in its unique historical place. He distinguishes them in terms of their economic mode of life, but also in terms of the stage of development of the human being that they represent. 'The orient'—that is, the ancient near east—was a tranquil pastoral society, grounded on submission to patriarchal authority and childlike religious feeling.[9] Ancient Egypt was an agricultural society, which developed landed property, civil administration, and the practical arts.[10] The Phoenicians, in contrast to the Egyptians, were oriented outward, toward the sea and toward trade between peoples.[11] The achievements of both were taken up, but also entirely transformed, by the Greeks, in whom there first arose a new sense of beauty and a love of freedom.[12] With the Romans, we see a pride in conquering one's desires

[8] Herder, xiii. 319, *Ideas*, Book VIII, ch. 4.
[9] Herder, v. 477–80; Forster, 273–6.
[10] Herder, v. 489–92; Forster, 280–4.
[11] Herder, v. 493–4; Forster, 284–5.
[12] Herder, v. 494–9; Forster, 285–9.

and sensual gratification, devotion to the fatherland, and the aspiration to build an entire world.[13] Herder also compares the successive ancient cultures to stages in the development of a single human being. The Orient is the infancy of the species, Egypt its boyhood, Greece its youth, Rome its coming to manhood. (In these terms, he significantly depicts his own century as an old man, a point to which we will return later.)

Each way of thinking represents a distinct side of humanity, and seeks a kind of happiness distinct from the others.[14] For this reason, Herder observes that every people regards the way of thinking of all the others with intolerance. The agricultural Egyptian despises both the Oriental shepherd and the Phoenician seafarer; the Greek regards all other peoples as barbarians.[15] Yet 'prejudice is good in its time, since it renders happy. It forces peoples together into their center, makes them firmer on their tribal stem, more blooming in their kind'.[16]

There is one serious danger here, of which Herder is aware, and which he takes steps to avoid. We might be tempted to ascribe to him the view that ways of thinking and ways of being happy are essentially cultural and collective, rather than individual. But this is the exact opposite of what he holds. 'No one in the world feels the weakness of general characterizing more than I,' he declares.[17] All thinking, and feeling, and the constitution of any human happiness, are always at bottom expressions of the distinctive individuality of particular human beings. In some ages there is greater uniformity, greater dependency on social authority or conformity to collective standards than in others (more in Oriental or Egyptian than in Greek culture, and more in all of these than in modern culture). But, in every culture, Herder holds, there is more individual variation in thinking and feeling than there is cultural uniformity. This is truest of all when it comes to what is most fundamental to human life, the conditions for happiness. As he says later in the *Ideas*: 'Happiness is an internal state; and therefore its standard is not seated without us, but in the breast of every individual, where alone it can be determined'.[18]

Herder's interest in generalizing about ages and peoples, and distinguishing different cultural ways of thinking and feeling, revolve around two main points. The first is the radical differences between their ways

[13] Herder, v. 499–501; Forster, 289–91.
[14] Herder, v. 508–9; Forster, 296.
[15] Herder, v. 487; Forster, 281.
[16] Herder, v. 510; Forster, 297.
[17] Herder, v. 501; Forster, 291.
[18] Herder, xiii. 333; *Ideas*, Book VIII, ch. 5.

of thinking and consequently of the conditions under which life—and a flourishing and happy life—is possible for each. These give rise to contrasting conceptions of morality and virtue that differ from one another just as radically as the mode of life and the happiness possible to each. The second—which Herder refers to at one point as 'my great theme'—is the historical continuity between the different ages, the necessary development from each stage of history to the next, which, Herder claims, displays the plan of Providence.[19]

The Failings of the Present Age

Herder views his own age as approaching its own history in an entirely wrong way, because it applies its own way of thinking, its own values, its own conceptions of happiness and virtue—in a word, what he calls (using a typical Enlightenment term) its own 'prejudices'—to other ages—for which they provide the wrong measure and result only in uncomprehending hostility. The present age, for instance, identifies the so-called 'despotism' of the ancient Orient or the hierarchy (priestcraft) of the Egyptians with the closest approximations to them in our own time, and because it rightly rejects these in relation to our happiness and our way of thinking, it fails to see how they were both right and necessary for the time in which they occurred.[20] It is, he says, as if an old man expected a baby, or a young boy, to live as he does, to think the same things, to value the same things, to take pleasure in the same things, as if we expected a baby or child or youth to be happy living a life suited to an old man.[21]

Herder also accepts the traditional idea that each people, and each stage of history, also goes through something like a life-cycle, involving a period of immaturity and growth, then a period of decline and decay, between them a time of flourishing, during which alone it fully enjoys the unique happiness proper to it.[22] In these cycles, peoples, nations, and whole ways of life and ways of thinking arise, mature, and then fade away to make room for others. This leads to the other main point—Herder's 'great theme'—that in the course of history a providential plan becomes visible,

[19] Herder, v. 511; Forster, 298–9. [20] Herder, v. 490; Forster, 282.
[21] Herder, v. 486, 489–90; Forster, 279, 282–3. [22] Herder, v. 509; Forster, 296.

a necessary narrative structure of history, so that through all its changes 'humanity ever remains only humanity'.[23]

According to Herder, the present age misjudges other ages on this score as well, by failing to appreciate how each earlier stage of history was necessary not only for its own time but also for the entire process, and in that sense also necessary for the present age itself. Thus he accuses his contemporaries not only of condemning other ages by judging them according to the wrong standards, but also of failing to appreciate with gratitude how past ages were necessary to make possible their own way of thinking and its happiness. He insists, however, that it is not his aim to defend the way of thinking of other ages but only to explain them.[24] Both praise and blame of the institutions of other times, he thinks, too often overlooks that what seems to us good and what seems to us bad are, in relation to the circumstances and way of thinking of that age, necessary to each other—as the slavery of the Helots was necessary for Spartan virtue,[25] and as periods of darkness were necessary to make our own enlightened age possible.[26]

The claims about necessity here are important to Herder, and he holds them in a very strong form. Later in the *Ideas* he states as his 'principal law': 'Everywhere on our earth whatever could be, has been, according to the situation and wants of the place, the circumstances and occasions of the times, and the native or generated character of the people'.[27] This 'law' would seem to be for Herder—to put it in Kantian terms—a regulative idea for the 'explanation' he seeks. We would fully explain a people or an age when we see how its feelings and thoughts, its actions and values, its strivings and its happiness, are all necessary to one another, and also necessary not only to the people's geographical and economic circumstances but also to its place in the historical succession of ages through which human nature develops.

We can best understand Herder's critique of his own age if we view it as simply an application to the present age of the basic principles of his theory of history. The present age, like every age, is intolerant of others, and regards its way of thinking and its happiness as the only way of thinking, the only way to be happy. It is this thought, I suggest, rather than anything

[23] Herder, v. 511; Forster, 298.
[24] Herder, v. 526; Forster, 309.
[25] Herder, v. 508; Forster, 295.
[26] Herder, v. 525–6; Forster, 308–9.
[27] Herder, xiv. 83; *Ideas*, Book XII, ch. 6.

Herder found in the thoughts or writings of Enlightenment thinkers, that leads him to depict the Enlightenment as viewing itself with arrogant complacency as the final goal of history.[28] The other main point to notice is that in relation to his conception of the life-cycle of an age, Herder regards his own time as a period of decline and (as he repeatedly says) of 'exhaustion'.[29] This is the deeper meaning of his metaphorical description of the present age as an old man. It colors his interpretation of it as a one-sidedly philosophical age, characterized by hyper-intellectualism, the separation of head and heart, a loss of faith in religion and of commitment to fatherland, and, in sum, what he calls its 'mediocrity of soul'[30] and its 'human misery'.[31] For Herder the high point of modernity was the practical flourishing of mechanical inventions in the seventeenth century, which formed the age's models of science, its politics, and its philosophy.[32] The high point politically came with the reign of Louis XIV.[33] The deepest mistake of the age is to think of its philosophy, its knowledge, its self-understanding, as the culmination of all history, since in Herder's view these are not a culmination of anything but merely symptoms of decadence.

This also accounts for the fact that Herder ends his manifesto with an enthusiastic expression of hope for a new age, in which (as in all ages past), the achievements of the present age will be appropriated and at the same time transformed into something entirely new and higher. It is in this spirit that Herder praises Enlightenment ideals—knowledge, freedom, equality, sociability.[34] In their present form, to be sure, they often do as much harm as good, leading to shallowness, corruption, and misery. Their true meaning and value will come to be appreciated only in the future, when they will be seen as mere fragments of a larger whole, as what he calls 'the fragment of life' that we presently are, after the plan of Providence eventually reveals them as tools of a larger purpose than we are now in a position to understand.[35] Herder's critique of his own age is simply a consequence of his acceptance of a form of historical optimism, of an objective natural teleology of history.

[28] Herder, v. 559–60; Forster, 334–5.
[29] Herder, v. 538, 556, 582; Forster, 319, 333, 355.
[30] Herder, v. 583; Forster, 356.
[31] Herder, v. 526, cf. v. 538, 550; Forster, 310, cf. 319, 328.
[32] Herder, v. 536–7; Forster, 317–18.
[33] Herder, v. 581; Forster, 354.
[34] Herder, v. 575; Forster, 349.
[35] Herder, v. 584–6, cf. v. 513; Forster, 356–8, cf. 299.

Cultural Relativism

Herder's view of history and society, as I have already mentioned, has often been described as 'cultural relativism'. This term is used in such varied, imprecise, and even confused senses that until we make it more precise, it would be just as pointless to deny that it describes Herder's views as it is uninformative to apply it to them. But we should realize that such terms were not current in Herder's day and he never describes his own views using them. When we represent Herder's views using terms like 'relativism' and 'historicism', there is a serious danger that we will not only distort them but also miss parts of them that might help us decide what is right and wrong with the later views bearing such names.

'Relativism', like many philosophical theses, but to a far greater extent than most, finds itself perched on a thin ledge between, on the one side, trivial truth, and on the other, absurd falsity, self-contradictoriness, or sheer nonsense. Answering even the simplest questions often forces a relativist to choose between the one abyss or the other. Herder's enthusiastic and undisciplined style of thinking and writing often opens him to charges of inconsistency, but I think there are some things we can note about his views that help exonerate him from the charges of incoherence or self-refutation that are often brought against various forms of relativism.

First, although Herder is aware how difficult it is to understand other times and other cultures, he never doubts that the kinds of claims he means to make about ancient societies, as well as about his own age, are *objectively true*. There is for him only one objectively correct understanding of any age—namely, the one suited uniquely to that age itself, the necessity of its relation to its geographical and historical conditions, and the necessary connection of its own way of thinking and acting. Any other understanding of it, such as one drawn from the standards and prejudices of a different age, is objectively false. Historical or cultural truth is not at all relative to the perspective of the observer, although our perspective may make that objective truth more difficult to know.

Second, Herder also rejects the 'cultural determinism' that belongs to some more recent anthropologists such as Ruth Benedict or Clifford Geertz. This is the view that some entity called a *culture* decisively conditions all individual thinking and provides the only criterion for the correctness of

the thoughts and actions of individuals in the culture. On the contrary, for Herder the deepest truth lies in individuality, and the true measure of the happiness and virtue of individuals is always their individual life. Culture provides only the necessary context for that endless richness.

Herder's closest approach to what we now call relativism is his view that the character of human happiness, and consequently also of morality and virtue, varies fundamentally from culture to culture and age to age. But this view is best seen as a consequence of combining two other views:

1. *Moral Eudaimonism* (ME): valid moral standards depend on the nature and means to the happiness of those to whom they apply.
2. *Variability of Happiness* (VH): there is no invariant human nature that determines what is best for all human beings. The content of human happiness varies greatly culture to culture, and even more from individual to individual.

ME is held by many philosophers, including Aristotle and the utilitarians—though eudaimonists certainly disagree among themselves about what happiness is and precisely how moral standards depend on it. VH is held by fewer philosophers, and seems to be rejected by Aristotle and most other ancient eudaimonistic schools of ethics. But one philosopher in Herder's time who very explicitly embraces VH is Kant. Kant, however, clearly rejects ME. Nevertheless, it would be absurd to regard Aristotle or Bentham, or Kant, as moral or cultural relativists. And it would seem strange that Herder should become a relativist merely by combining two already familiar views, neither of which is naturally thought of as relativist. ME proposes a universal and purely objective standard for determining the content of morality. VH tells us that this standard applies very differently to different cultures, ages, and individuals. If relativism is simply the view that the same objective standard will apply differently to different people when the facts of their situations differ, then it begins to look trivially true. If relativism is the thesis that the combination of ME and VH entail that there is no objective truth about happiness or virtue, then the essence of relativism seems to consist in nothing but an elementary falsehood resulting from an obvious non sequitur.

Yet this is not an unfamiliar dilemma for moral or cultural relativism, which often results from mistaking the practically important but relatively superficial modes of valuation involved in local customs for the kinds of

fundamental principles and values that might both ground different customs and rationally account for variations among them. William Graham Sumner, for example, most often states cultural relativism as the doctrine that the moral rightness of acts for members of a society is determined by the society's beliefs about what is morally right. At times, however, he claims that acts or customs are morally right because they are well adapted to the needs and circumstances of people, and at those points he seems to treat the moral beliefs of a society as merely an empirically reliable guide to what has the property of being well adapted.[36] Similar shifts occur in the writings of other cultural relativists, such as Melville Herskovits.[37]

That moral rightness consists in adaptation to a certain way of life seems quite close to Herder's view, but it is doubtful whether it by itself deserves the name 'relativism'. On the contrary, it looks more like a still inarticulate expression of the sort of basic value that might ground the variation among human practices when it is applied to very different circumstances. Universalism about fundamental values or principles obviously becomes more and not less plausible if you assume radical differences among people and cultures, since it is easier to account for radical differences in cultural practices on universal principles when these principles are being applied to very different things. It is only if we assume that human nature and the conditions of human life are relatively uniform across cultures that radical differences in fundamental values or principles would be required to explain the fact that people's customs differ greatly.

Herder's closest approach to relativism seems to me to consist not so much in any view that he holds as in his awareness of a certain *problem* that arises when we become aware of the radical variation in historical and cultural ways of thinking. Herder perhaps states this problem most clearly in an early fragment to which his German editors gave the (misleading) title *Change of Taste* (1766). There Herder observes that as soon as we become convinced on the basis of reasons that anything is true or good or beautiful we straight away expect that everyone else will agree with us on the basis of those same reasons. But in fact we find (clearly this is one surprising result of Herder's approach to our understanding of alien ages and cultures) that others may regard the same things as false, bad, or ugly, also on the

[36] Sumner, 'Folkways', in John Ladd (ed.), *Ethical Relativism* (Belmont, Calif.: Wadsworth, 1973), 23–9.

[37] Herskovits, 'Cultural Relativism and Cultural Values', ibid., 58–78.

basis of reasons, so that (as Herder puts it) 'truth, beauty and moral value is a phantom that appears to each person in another way, in another shape, a true *Proteus* who by means of a magic mirror ever changes and never shows himself as the same'.[38]

Herder realizes that this 'contradiction', as he calls it, may lead us to doubt our own convictions, and even tempt us to become principled skeptics about everything. But he firmly rejects this reaction, both in *Change of Taste* and when the analogous problem arises in *This Too a Philosophy of History*.[39] When in *Change of Taste* Herder asks the question 'Is not truth, fairness and moral goodness the same at all times?' he replies unhesitatingly in the affirmative, though he significantly admits that it is impossible to be entirely comfortable in answering this way.[40] And in other places, such as a journal entry dated three years later, he flirts with the contrary answer.[41] Quite often, I submit, 'relativism' is a name given to any pattern of thinking characterized by an awareness of the very real problem Herder is raising in *Change of Taste*, followed rapidly by a facile, complacent (and fundamentally dishonest) dismissal of the problem by offering oneself some set of skeptical sounding assertions that, when more closely examined, simply don't make coherent sense. (If we want a precise definition of 'cultural relativism', that would be mine.) Herder is a cultural relativist to the extent that he acknowledges the problem, but I do not think he can be accused of its facile and incoherent dismissal, so he is only half-relativist (the honest and correct half, that leaves us in a state of intellectual dissatisfaction, not the dishonest and complacent half, that tries to run away from that state).

However, Herder did eventually formulate a solution to the problem, at least in regard to the variation involving different historical ages. It was not a facile solution, and, to his credit, it was also one with which he was never wholly satisfied. This is his theory of the necessary teleological development of humanity through historical stages. We reconcile the way of thinking that characterizes earlier ages with our own way of thinking by understanding those earlier ages as necessary not only to the place in history that we occupy, but also to the future development of humanity in accordance with a plan of Providence. In his historical manifesto of 1774, Herder's position seems to be that we are capable of discerning that there

[38] VdG, i. 149, Forster, 247.
[39] Herder, v. 583; Forster, 355, cf. VdG, i. 151, Forster, 248–9.
[40] VdG, i. 160, Forster, 256.
[41] *Werke*, iv. 472.

is such a plan, but we realize that its end or goal is forever hidden from us, because the further development of the plan will take humanity beyond us into a future we cannot pretend to understand.[42] In the *Ideas*, Herder has found a name for the goal of this process—he calls it *Humanität*—a term he defines as 'reason and equity in all conditions and all occupations of human beings'.[43] In other words, the resolution to the problem of historical variation is that the human species, through the secret plan of Providence, is striving ever forward toward precisely those ideals that were central to the Enlightenment.

One claim commonly ascribed to Herder is that different cultures or ages are 'incommensurable'.[44] This way of putting it seems motivated mainly by a desire to move Herder's views in the direction of currently more familiar ones, thus reflecting either the interpreter's wishes or displaying his limited philosophical imagination. Herder thinks each culture and time should be explained from within, as a unique development of organic forces rather than as an instance of universal natural laws like those of Newtonian mechanics. He likewise rejects any standards of evaluation that supposedly rest on a uniform and invariant human nature. Hence correct standards of happiness and virtue are not 'commensurable' in *that* way. But the whole point of Herder's philosophy of history is to enable us to construct a narrative progression leading from earlier times to our own, in which each culture and each age are understood in their own terms and yet also related to the progression of human history under the idea of *Humanität*. The latter relation is not one of incommensurability but rather of a new kind of distinctively *historical* commensurability.

Herder's counter-Enlightenment admirers are usually aware of this, but that doesn't mean they have to like it. So when they come to the topic of *Humanität*, their stance is often to grumble, prevaricate, and condescend. 'Perhaps,' admits Isaiah Berlin, 'Herder did come to believe [in *Humanität* as a single uniting goal of history], or at least to believe he believed it.'[45] When it comes to the central unifying idea of Herder's mature philosophy of history, it seems that I differ from Berlin only in believing that Herder

[42] Herder, v. 513; Forster, 299. [43] Herder, xiv. 230; *Ideas*, Book XV, ch. 2.

[44] For example, see Isaiah Berlin, *Three Critics of the Enlightenment: Vico, Hamann, Herder*, ed. Henry Hardy (Princeton, NJ: Princeton University Press, 1997), 234 ff.; Frederick Beiser, *The Fate of Reason* (Cambridge, Mass.: Harvard University Press, 1987), 143–4.

[45] Berlin, *Three Critics of the Enlightenment*, 234.

believed what Herder believed he believed. Or, according to Michael Forster, Herder had doubts about *Humanität*, 'just below the surface'.[46] Or maybe he had no such doubts after all, but we can still ignore *Humanitat*, simply because *we* have them:

> [Herder's] philosophy of history is initially likely to seem striking and interesting mainly for its development of a teleological conception of history as the progressive realization of 'reason' and 'humanity'—a conception which anticipated and strongly influenced Hegel, among others. However, this conception is highly dubious on reflection, and is arguably *not* one of Herder's main achievements in this area.[47]

My argument in this paper, however, is that Herder's historical teleology is integral to his philosophy of history, hence also 'arguably' inseparable from his main achievements in that area. (Isn't it strange how the word 'arguably' permits people to make directly contradictory claims, *arguably* with equal justification?) Along with Kant, Herder is the chief source for the great nineteenth-century theories of history in the German idealist tradition: those of Fichte, Schelling, Hegel, and Marx. This is an intellectual tradition for which no one needs to apologize.

Kant and Herder

Herder and Kant did disagree about many things. Herder didn't think there was such a thing as a priori knowledge, though, as it is with many empiricists (and especially pragmatists), this is more a matter of irresponsible emotive

[46] Forster, *Herder: Philosophical Writings*, Introduction, pp. xxvi ff. Taken literally, what Forster says here is entirely true, I think. If I am hinting at a criticism of it, that is because I hear in Forster's remark a note of grumbling, as though he wishes Herder had not made so much of *Humanität*. Maybe I am hearing something in the quoted remark that isn't intended by Forster, and not even really there. My own view is that Herder was quite right to think that a theory of history of the sort he was presenting requires some telos to the historical narrative, and he was showing the instincts of a good *Aufklärer* in choosing *Humanitat* as this telos. In fact, he was being a better example of the moralizing *Aufklärung* than was Kant, whose natural telos in history is not a moral ideal of any kind but only the open-ended development of the natural species-predispositions of humanity. I also think Herder showed both good sense and intellectual integrity in expressing occasional doubts that history has this (or any other) telos. The best quality in any philosopher is to think through systematically and consequently what his theory commits him to, to draw these conclusions resolutely, and then to express the doubts about them that any sensible person feels when faced with any answer to philosophical questions, which always leave us baffled no matter how long or how well we think about them.

[47] Forster, *SEP*.

gesticulating than well-thought-out epistemology. Herder, like many in the following generation of German philosophers, rejected Kant's division of our theoretical faculties into sense and understanding, and of our faculty of desire into reason and inclination. In matters of taste, he rejected Kant's classicism, and in metaphysics he favored an organism composed eclectically out of doctrines from Spinoza and Leibniz that Kant regarded as enthusiasm.

In the philosophy of history, however, we have a fairly definite account of their differences in the form of the criticisms of Kant that Herder offered in the *Ideas*, and to which Kant replied in his review of the second volume of that work. These differences are real, but I do not think they are as fundamental as some would maintain they are. In fact, I think we will see that all the main disagreements concern the precise way in which we should understand the objective natural teleology of history—a teleology both thinkers equally accepted, though Herder perhaps gave it a more metaphysical, theological, and constitutive interpretation, while for Kant this teleology was always a regulative principle of reason in service of maximizing the comprehensibility of history to us. The principal *difference* between Herder and Kant, I would say, is that Kant, like many in his age, never fully comprehended or tried to practice the difficult method of understanding other ages that Herder advocated, outlined, and tried to execute. For the same reason, he never truly confronted the 'change of taste' problem to which Herder's theory of Providence's plan of development toward *Humanität* offers, even in Herder's own view, I think, only a partial and less than fully satisfactory solution. But these most important differences are not disagreements so much as merely omissions on Kant's part to take up the most original and challenging ideas of his erstwhile student.

When it comes to actual disagreements in the philosophy of history, I suggest that the main ones are three in number.

1. The role of the political state in the teleology of history.
2. The place of human happiness among the ends of nature (or Providence).
3. The understanding of earlier ages—especially the unhappiness found in them—as means to later stages in historical development.

In *Idea for a Universal History with a Cosmopolitan Aim* (1784), Kant argues that reason requires us to seek for an unconscious natural purposiveness

in human history as a way of maximizing its theoretical intelligibility. The purposiveness he posits as a regulative principle is the indefinite development of the faculties of the species. This development, he argues, is seen empirically to occur through the mechanism of the unsociable sociability of human nature, which leads people to compete with one another and by means of this competition progressively to enrich their faculties. The social antagonism that serves as nature's means, however, when it manifests itself in the form of violence among human beings, eventually threatens nature's end. The end both of nature and of moral reason therefore requires the establishment of a law-governed civil society capable of enforcing peace with justice among human beings. Nature's end is thus served by the tendency in history for human beings to perfect the form of a civil constitution administering justice, and this, in turn, Kant argues, will require establishing peaceful relations between states through a federation among them. In the course of explaining the difficulty of seeking a just constitution, Kant declares that the human being, in the social condition, is 'an animal who needs a master', yet the problem is that this master will also be an animal who needs a master, making the construction of a perfectly just constitution a task with no ideally perfect practicable solution.[48]

In the second volume of the *Ideas*, Herder attacks Kant's proposition that 'the human being is an animal who needs a master' as an 'easy but evil principle'. Kant replies that it is easy because everywhere confirmed by sad experience, but he denies it is evil.[49] Kant does not mean, of course, that human beings are born for slavery, since the whole point is that they require a coercively enforced law to protect the external freedom that goes with their dignity as rational beings. The real disagreement concerns the role of the political state in the natural (or Providential) teleology of history. Kant regards coercion as an essential and permanent part of human life, owing to the unsociable sociability and hence the innate viciousness of human nature. Herder views the progress of humanity as resting on the victory of nobler and milder human predispositions, characterized by *Humanität*. We might see Herder as hoping for the eventual abolition or withering away of the state, whereas Kant entertains no such hope. This gap narrows, however, when, in the *Religion* Kant comes to see the highest

[48] *IAG*, viii. 23. [49] *RH*, viii. 64.

manifestation of historical teleology as taking place not in the coercive state but in the 'moral commonwealth'—whose model is the religious community or church.

The second disagreement arises out of Kant's view that the natural teleology of history makes use not merely of social antagonism but also of human discontent as a spur to the development of our species faculties. For Kant, human happiness is an end of prudential and even of moral reason, but it is not an end of nature. Human beings were not put on the earth to be happy, their struggle to achieve happiness, though required by reason, is profoundly contrary to nature's plan for them. For Kant, the struggle to be happy is a struggle against nature.[50]

Herder agrees that perfect happiness or contentment will never be possible for human beings.[51] But for him happiness is not a delusive will-of-the wisp as it is for Kant, 'Every living being [says Herder] rejoices in his existence; he does not inquire, he does not strictly examine, why he exists: his existence is to him an end, and his end is existence'.[52] Herder regards Kant's notion that the plan of Providence would require human beings to be permanently *unhappy* as an impious indictment of the Deity. For Kant, on the contrary, there are higher ends than happiness, and the moral worth of our person, rather than the happiness (the worth of our state or condition) is the correct measure of the meaning and value of our existence. Thus, if human beings lived, as Kant imagines the inhabitants of Tahiti to do, merely for the pleasure of existing, without developing or exercising their properly human faculties, then Kant wonders 'whether it would not have been just as good to have this island populated with happy sheep and cattle as with human beings'.[53] To prevent misunderstanding, we need to add here that of course Kant does not believe that it would be just, or even possible, for Europeans to 'civilize' other peoples. With full Kantian severity, he condemns the unjust actions of European colonialism, in invading and exploiting the inhabitants of other parts of the world.[54] But he does think that human beings everywhere give meaning to their existence only by developing, according to their own lights, the capacities of the human species, thus fulfilling the purpose of both nature and reason.

[50] *Critique of the Power of Judgment*, v. 429–31.
[51] Herder, xiv. 333; *Ideas*, Book VIII, ch. 5.
[52] Herder, xiv. 337; *Ideas*, Book VIII, ch. 5.
[53] *RH*, viii. 65.
[54] *EF*, viii. 358–9; *MS*, vi. 266.

This issue is closely related to the third one. Herder finds absurd Kant's idea that the end of nature for the human species—namely, its indefinite development of its rational predispositions, should be thought of as fulfilled only in the open-ended future of the human species as a whole, never in the happiness of individual human beings at any given time.

> If someone said that not the individual human being but humankind is to be educated, then he speaks unintelligibly for me, since kind and species are only general concepts, except only insofar as they exist in individual beings.—It is as if I spoke of animality, minerality and metality in general and adorned them with the most splendid attributes, which, however, contradict one another in single individuals![55]

Kant's reply is that when he speaks of the end of nature being fulfilled only in the human species as a whole, he does not mean the abstract general concept of the species but rather something particular and concrete—'the *whole* of a series generations going (indeterminably) into the infinite'.[56] In other words, Kant thinks the end of nature in history, the endless development of our species capacities, is fulfilled in the same way that Herder thinks the end of nature is fulfilled in the human species through the endless progress toward *Humanität*. The disagreement is over whether it is a fitting way to think of the plan of Providence to regard the happiness (or rather, for Kant, mainly the discontent) of all individuals, and of every generation, as nature's means for fulfilling an end that no individual and no generation enjoys. You might say that Herder is afraid that in Kant's view of history God is guilty of violating the Kantian principle that humanity in everyone's person should be treated as an end, not merely as a means. 'Not a thing in the whole of God's realm', Herder declares, 'is *only* means—everything is *means* and *end* simultaneously'.[57] Of course, for Herder, the only way individual human beings can be an end is by being happy, rejoicing in their own existence, whereas for Kant, happiness is only a conditioned good, and the worth of human existence is measured by human perfection, especially the moral perfection of will. Yet Herder himself has a problem of the same kind, owing to his doctrine that there is a life cycle for every age and every people, with its true happiness attained only during the period of its maturity.[58] It is not clear how he can clear

[55] *Ideas*, xiii. 345–6; *RH*, viii. 65. [56] *RH*, viii. 65. [57] Herder, v. 527; Forster, 310.
[58] Herder, v. 511; Forster, 298.

himself of the charge that those who endure less happiness during periods of growth and decline are not being treated by Providence as mere means to the end of the happiness of those who enjoy the full flourishing of their age.

The whole issue, however, seems to matter only if you are preoccupied with the problem of theodicy. Once we begin to think of the human species in history as struggling to create meaning in the context of a nature that sometimes treats us kindly, more often cruelly (its final act toward each of us will be murder), but a nature that is always fundamentally blind and indifferent to human concerns, we no longer expect nature to obey the Categorical Imperative, and the point of controversy here between Herder and Kant just disappears for us.[59]

These real points of dispute between Kant and Herder are interesting and significant, but not nearly as important as the points of agreement that made them all possible. If our aim is to show Herder as the superior thinker, then we would be wise not to portray him as a critic of Kant or of the Enlightenment, but rather to focus on his positive accomplishments, which (however he may have understood them) in fact enabled the Enlightenment's values and vision to be extended to a richer appreciation of the full wealth of human life and history.

[59] For most of us, anyway. A few seem to think that just to look at history as progressive is to make oneself complicit in the evil of treating past generations as mere means to future ones. One who thinks this way is Hannah Arendt: 'In Kant himself there is this contradiction: Infinite progress is the law of the human species; at the same time, man's dignity demands that he be seen (every single one of us) in his particularity, and, as such, be seen—but without any comparison and independent of time—as reflecting mankind in general. In other words, the very idea of progress—if it is more than a change in circumstances and an improvement of the world—contradicts Kant's notion of man's dignity. It is against human dignity to believe in progress' (Hannah Arendt, *Lectures on Kant's Political Philosophy*, ed. Ronald Beiner (Chicago: University of Chicago Press, 1982), 77. Similar charges are found in William Galston, *Kant and the Problem of History* (Chicago: University of Chicago Press, 1975), 231–5; Paul Stern, 'The Problem of History and the Temporality of Kant's Ethics', *Review of Metaphysics*, 39 (1986), 535–9; and Louis Dupré, 'Kant's Theory of History and Progress', *Review of Metaphysics*, 52 (1998)). The real target of some of this is not the idea of progress but Kant's supposed view that our real selves are noumenal and atemporal. This is not the place to say why that reading of Kant is wrong. In our postmodern, post-historical, and post-human age, however, no doubt the very idea of progress is also suspect among those who want to represent themselves as among the more advanced thinkers of the time (and do not mind contradicting themselves in the process, since contradicting oneself always makes even the driest intellectual, oh, so mysterious and sexy, at least in his own eyes). It is even easier to contradict yourself if you develop the habit of seeing contradictions where there are none—as between the idea that the human species should strive to progress in history and the idea that its final moral end, the striving for which is necessary if it is to live up to its dignity as an end in itself, is to participate in this striving. To understand Kant's doctrines correctly, and still think there is a contradiction between them, is almost to think that, for all p, p must contradict p.

In his later writings (starting about 1789), Kant offers three 'maxims' for our use of reason (or understanding):

1. Think for yourself.
2. Think from the standpoint of everyone else.
3. Think consistently.

To me, these maxims exemplify the spirit of Kantian philosophy, and of the Enlightenment, as profoundly as any thoughts he ever expressed. I view Herder's great contribution as, in effect, a remarkably original expansion on the second rule. Kant always regarded all three rules as extremely difficult to satisfy, and impossible for flawed and limited creatures like ourselves ever to follow perfectly. Herder showed us a new dimension of difficulty in following the second rule of which Kant and most of his contemporaries had little or no conception, but he also opened up for us a new way of thinking about culture and history that make possible a new kind of striving to follow it.

Faith in Historical Progress as a Rational Faith

Herder and Kant both seek to comprehend history by supposing that historical events promote a certain end (or ends). The ends they use are different in the two cases, so their theories of history differ materially in certain respects. But I think the grounds, both theoretical and empirical, that they use to justify their theories are quite similar. For Herder, the end of history is *Humanität*—'reason and equity in all conditions and all occupations of human beings'. This goal emphasizes both the endless variety of human institutions, experiences of life, and ways of self-understanding, and also a tendency of different human beings to understand this variety, rationally and equitably, as a valued expression of different sides of human nature. For Kant, the basic tendency in history is the endless development of our species' predispositions through unsociable sociability, which can continue (after the coming of civilization) only through progress toward the idea of a law-governed civil society administering justice, and then (at a certain stage in the development of states) through progress toward an international organization of states maintaining a just peace between them. Neither philosopher maintains that progress toward these ends is

smooth or uninterrupted. Kant even holds that the idea of a law-governed civil society is to be employed by historians 'to show how far humanity has approached this final end in different ages, or how far removed it has been from it, and what is still to be done for its attainment'.[60] In other words, there is no dogmatic optimism about the course of human events.

In both Kant and Herder, a teleological view of history can be seen as a response to two simultaneous demands of rational inquiry: one theoretical, the other practical. On the one hand, as rational inquirers into human history, we seek some kind of systematic comprehension of the facts, by showing how they exhibit certain underlying tendencies. On the other, as historical agents, we seek to comprehend our own actions as part of an ongoing historical pattern in which we may contribute toward some worthwhile end, which history exhibited before our arrival and can be expected also to exhibit after we are gone. Both motives can be understood as rationally required if we are to make sense of history, whether as rational inquirers or as purposive agents. As rational inquirers, we seek for something to unify the events of history into a meaningful pattern. Kant argues in the *Critique of the Power of Judgment* that in looking for a systematic order among natural phenomena that cannot be explained by mechanistic causal laws, it is rationally required that we employ the concept of a natural end. This is what Kant does with his theory of the development of our species' predispositions through unsociable sociability, and what Herder does when he seeks a connectedness among different peoples and ages, ultimately employing the idea of *Humanität*. As agents, we seek to understand our own actions as contributing to a meaningful historical process. We seek for historical trends or tendencies that harmonize with our human strivings for justice, peace, equity, and reason, so that we may both understand our agency better and use it to contribute to something in history we consider worthwhile. These are not merely optional aims for us in so far as we are rational knowers and purposive agents who also recognize ourselves as historical beings. If we choose not to look for meaningful purposive patterns in history, and not to strive for a better future for our species in history, this is in effect to give up on our own rationality as historical beings.

[60] Ak, viii. 468.

Of course, for both purposes, we must be flexible—as both Kant and Herder are—in our conception of the ends we ascribe to history. Whatever ends we see in history, and whatever historical ends we set for ourselves, we must conceive them only provisionally, remaining open to the empirical data and to our own fallibility in the way we specify them. *Humanität* is a conception that might be specified in endlessly many ways, and it is our task both as knowers and agents to adapt our conception of it to fit our knowledge both of what has happened (and is happening) in history and of what ends are worthwhile for human beings to pursue in history. Likewise, the content of human predispositions is not something fixed in advance, just as the idea of a wholly just civil society is something constantly evolving as we learn more about human social life and about the requirements of right itself. But without some conception of this sort, as historical inquirers we would have to abandon the task of making systematic sense of history, and as agents we would have to cease to think of ourselves as part of a historical progression at all. Moreover, if we undertake both the theoretical and the practical tasks of making rational sense of history, we are bound to try to harmonize the two projects, so that the sense history makes to us as agents fits into the meaning we comprehend in history as knowers.

We may, of course, choose to despair of the theoretical project of comprehending history as well as the practical project of contributing to a collective project of the human species through time. It would no doubt be possible simply to give up on both projects, based on the kinds of horrors and disappointments that have disillusioned many historical optimists during the twentieth century. This would seem to be precisely what has been done by those who distance themselves from the entire Enlightenment conception of historical progress. The crucial point to make, however, is that no set of empirical facts could actually justify such an act of despair. It can be made sense of only as a certain irrational emotional decision, prompted perhaps by the failure of certain specific (and always revisable) conceptions about where history is and ought to be heading. Instead of revising these conceptions when the facts indicated the necessity of so doing, some people found this task too hard or too emotionally costly, perhaps owing to their irrational attachment to these provisional conceptions. So, rather than give them up, they chose to abandon the entire rational enterprise of understanding history and orienting one's actions toward it. In contrast to this, Kant, Herder, and the entire Enlightenment tradition of thinking

about history insist that we can rationally abandon neither the theoretical enterprise of comprehending human history nor the practical enterprise of contributing to a better future for humanity. This tradition insists instead on a sober perseverance in these enterprises, and an unwillingness to give up either.

This kind of situation, and the dilemmas it poses, is illuminated by thinking about it in terms of Robert Adams's conception of moral faith. For Adams, 'faith' is the name for a response to a situation in which we believe something that is closely bound up with a project of some kind to which we are committed, but where there are also grounds for a reasonable person to doubt these same beliefs. 'Faith' is the response that remains true to the project by holding to the beliefs it presupposes, in spite of the temptations to doubt these beliefs and give up.[61] That faith occurs only where some degree of doubt is reasonable is the reason that Paul Tillich, for instance, claims (perhaps with an air of paradox) that doubt is a necessary element of faith. It is also the point being made by certain religious critics of dogmatic fundamentalism when they say: 'The opposite of faith is not doubt but certainty.' The Enlightenment conception of history, as exemplified in Kant and Herder, is an attempt to make sense of human history in the face of obvious reasons to doubt that it makes sense. Thus their philosophies of history exhibit *faith* in precisely Adams's sense.

Adams points out that faith, properly speaking, is not directed merely to propositions, or forms of words, but 'a stance in relation to something larger'.[62] In this discussion, Adams is discussing faith in *morality* and in moral ends. Religious people have faith in God, or in God's goodness, his promises, his love, or his mercy. Enlightenment historical faith is faith in human reason, as capable of comprehending our history as something meaningful, and in ourselves as rational beings, as capable of building a better human future. This Enlightenment faith can be a religious faith—both Kant and Herder relate their philosophy of history to the idea of divine providence. But it can also be a purely secular faith, and faith itself (as Adams describes it) need not have anything religious about it at all.

Adams notes that faith in something—in morality, or in divine purposes—is entirely compatible with being open to the evidence to revising

[61] See Robert M. Adams, *Finite and Infinite Goods* (Oxford: Oxford University Press, 1999), 373–89 (cited hereafter as 'Adams', by page number).

[62] Adams, 374.

your beliefs about what these are.[63] We have just noted that Herder and Kant both conceive the ends of history in such a way that they invite such empirically sensitive and revisionary attitudes, so it is a caricature of Enlightenment views of history to see them as blind to the facts or 'aprioristic' in some bad sense of the term. More generally, Adams rightly emphasizes that faith need not be an unreasoned belief, still less an irrational one.[64] In fact, he describes it in Aristotelian terms as the virtue that lies between the vices of credulity and incredulity.[65] I think this is true of faith when it is a virtue, though I think faith is not always a virtue—for we can misplace our faith by persisting in projects we should abandon and holding beliefs we should give up. In fact, I think faith remains a virtue only when it is fully rational. I would insist that faith, whenever it is a justified attitude, must be fully consistent with a moral principle that I agree with, which is often called 'evidentialism': the principle that our beliefs must always be proportioned to reasons or evidence. Some people use the word 'faith' to refer only to beliefs that violate the evidentialist principle—as in Mark Twain's wry definition: 'faith is when you believe something you know damn well ain't true.' But sometimes we have reason to doubt our beliefs but even better reason to stand by them, despite our temptation to give up on them. To do this is precisely to exhibit *faith* in Adams's sense. Faith, whenever it is rational and a virtue, is therefore even one of the demands the evidentialist principle makes on us.

Thus, in so far as faith is the virtue lying between incredulity and credulity, it involves beliefs we have some reason to doubt, but also reason to stand by, as part of our commitment to a project (theoretical, practical, moral) of some kind. Adams points out that not all the beliefs we hold are to be thought of as hypotheses to be tested and perhaps falsified.[66] This is especially true in the case of what Kant calls 'regulative principles of reason', or beliefs involved in the pursuit of ends we are rationally required to pursue. The project of making rational sense of the world, and the regulative principles involved in it, are not subject to empirical falsification. Neither is the project of giving meaning through our actions to our own life or the collective life of humanity, or, therefore, the belief that this is possible for us. I suspect that Adams may think that there is some refutation (or at least qualification) of evidentialism hidden in this last point. But I do

[63] Adams, 383–4. [64] Adams, 375. [65] Adams, 374. [66] Adams, 382–3.

not. Evidentialism, as I interpret it, tells us to be responsive to reasons and evidence in forming, maintaining and revising our beliefs. The regulative status of some of these beliefs, or their involvement in practical projects we should not abandon, seems to me merely one species of these reasons.

Faith, when it is a virtue, never has anything to do with wishful thinking—with the sad, or comical, or dangerous, human tendency to believe something because we want it to be true or because believing it is true makes us feel good. Wishful thinking is always an irrational way of forming beliefs, showing disrespect for ourselves as rational beings, and the beliefs formed in this way are usually erroneous and often dangerous. Of course, if we set ourselves some end, and hold the belief that our end is possible of attainment, then we probably also want it to be true that our end is possible. But here the direction of our motivation is just the opposite of that involved in wishful thinking. Wishful thinking believes something because it pleases us to believe it. In rational faith directed to a rational end, we believe our end is possible because we have rationally chosen to pursue it. Having this belief might also please us, but this pleasure is only an incidental result or by-product of the rational belief, not its cause. (If it is really the cause, then we are talking about something other, and less respectable, than rational faith.)

Wishful thinking, however, is far from being the only pattern of irrational belief formation. People are also subject to what we can call 'fearful thinking'. Their dread of an outcome sometimes makes them find it more likely, and this sometimes leads them into irrational patterns of behavior as they become obsessed with dangers that have been exaggerated by their imagination. Then they flounder about among emotionally charged ways of coping with the dangers. The dangers may be real, but in their fear they have lost their rational grip on how to deal with them. Faith is precisely the proper rational response to fearful thinking. This makes faith, properly placed, entirely rational and even an indispensable resource of human rationality.

Fearful thinking, for example, has had a fateful role to play in the American reaction to the events of 11 September 2001. It has made people think that Islamic terrorism is a greater threat to them than it really is. It led some people to think that an unprovoked military conquest of Iraq might be the only way of preventing weapons of mass destruction from falling into the hands of terrorists and being used against US cities. It caused

Americans to abandon their commitment to principles of human rights, to engage in the indefinite detention of certain individuals without charge or trial, and in acts of torture. Fearful thinking caused Americans to lose faith in themselves, in principles of international law, and due process of law. The same course of events also shows that when people lose faith in what they should have faith in they often replace this lost faith with a misplaced faith in something they should not have faith in. For when the fearful events of 11 September caused Americans to lose faith in what is decent in themselves it led them to put their faith in the wrong things—in acts of military aggression, vengeance, and injustice, and in a corrupt and incompetent political regime, from whose misdeeds we are now suffering, and from which the world is going to suffer for many years to come.

The Enlightenment conception of history, as exemplified in different ways by Kant and by Herder, is in this sense the result of an act of rational faith. Those who have rejected it on the basis of twentieth-century events such as the First World War, the Holocaust, or the failure and collapse of Eastern European communism, exhibit the kind of historical despair to which Enlightenment faith is exactly the needed response. Religious objections to Enlightenment optimism, for example, are often in large part expressions of despair over modernity, or humanity, or reason. To the extent that their God is the *totaliter aliter* of all these, it has to be reckoned a superstitious idol, or even an evil god. Some postmodernists are disillusioned Marxists, who would have done better to have kept faith with Marxism, while perhaps revising some outdated aims and brittle doctrines to fit reality, instead of letting historical despair be their last act of stubborn allegiance to an irrational dogmatism. Others are simply followers of an aesthetic fashion of despair, nihilism, and ironic detachment from all serious purposes in human life. In contrast, Enlightenment faith refuses to let the difficulty of making sense of history, or the painful frustrations caused by our disappointed historical hopes, push us into the despair of giving up on human history as a subject-matter for rational comprehension or on ourselves as historical agents.[67]

[67] People who despair of the Enlightenment view of history need not despair of everything, of course, as many of them would be eager to point out. But they are despairing of rational comprehension of a human endeavor considered as a whole, and of thinking of themselves as rational participants in it. Often people who do this try to portray it as merely a form of epistemic modesty, as though philosophical reflection on history were all by itself a kind of metaphysical extravagance. And they

Kant's sober insistence that we must look for rational purposiveness in history, and for a set of moral tasks arising out of our condition as historical beings, is a form of rational faith. Herder, too, as I have argued, did not repudiate the Enlightenment in this respect, but in the end kept faith with it. His greatest contributions were instead towards deepening and enriching it—in ways of which, like the radical aims of the Enlightenment themselves, we still cannot see the end.

might try to portray their denial of it as mere modesty and sobriety rather than as a despair of anything. I hope they will forgive me for not being convinced by such flattering self-portrayals. For once an intellectual project, such as the Enlightenment attempt to comprehend history (and its various nineteenth-century descendants—which all remain children of the Enlightenment, however much they may self-deceptively try to think of themselves as superior to it or as 'going beyond' it), there is simply no way of turning your back on it without in effect despairing of it, and of the functions of both theoretical and practical reason that it represents.

10

Moral Obligations and Social Commands[1]

SUSAN WOLF

In ordinary discourse, we sometimes use the language of right and wrong morally to evaluate actions. We talk about actions being morally required or obligatory, others as permissible, and still others as forbidden or wrong. On other occasions, we use the vocabulary of good and bad. In some moral theories and in much ordinary conversation, the difference between these sets of terms goes unnoticed—but there *is* a difference which is easily recognized when we are asked to attend to it. To say that an action is good is not the same as saying that it is obligatory. An action or type of action may be encouraged or praised without being morally required. Similarly, we may judge an action to be morally bad without finding it strictly immoral or wrong. We may discourage an action or criticize it without regarding it as forbidden.

Recognizing a difference between these sets of terms, however, does not amount to understanding it. The idea of a moral obligation or requirement, as opposed to that of an action that is (merely?) morally good, is especially puzzling, if not problematic.[2] It is the aim of this paper to understand it better.

[1] I owe thanks to the Mellon Foundation and to Oxford University's Centre for Ethics and Philosophy of Law for providing me with the most favorable circumstances possible in which to write this paper, and to audiences at Oxford, the Australian National University, the University of Melbourne, and the University of Sydney for very helpful discussions of an earlier draft.

[2] It may be noted that I do not in this paper contrast the obligatory with the *supererogatory*, which philosophers usually understand to refer to what *is above and beyond* the call of duty. In that understanding, to act in a way that is supererogatory is to do *more* or *better* than is morally required. The contrast I want to discuss here is not quite the same, for reasons that I hope will become clear as the paper develops.

Robert Adams has written with exceptional clarity and insight about the concept of moral obligation, arguing 'that a theory according to which moral obligation is constituted by divine commands . . . is the best theory on the subject for theists'.[3] There is reason, however, to hope for a theory of moral obligation that could be accepted by theists and non-theists alike. In what follows, I shall consider what seem to me to be the most popular as well as the most plausible theories of moral obligation, bringing out some of the difficulties as well as some of the advantages of each. As will be seen, my understanding of the concept of moral obligation shares much with Adams's conception. Like Adams, I think the nature of moral obligation is best understood against a background of an independently available conception of moral goodness and of good moral reasons, and, like him, I think central features of the notion of moral obligation can only be captured by a theory according to which obligations arise from actual social requirements. None the less, the account I shall be defending is a secular one, according to which obligations arise not from divine commands, but from human ones. Moreover, although I shall be defending what might be called a Social Command Theory of obligation, I defend it only to a point. As I shall argue, we use the language of obligation for a variety of purposes that are not all optimally served by the same understanding of the term. Disentangling these purposes and recognizing the different conceptions of obligation that best serve them will, I hope, shed light on the role the concept (or concepts) of obligation play(s) in moral thought.

Command Theories of Obligation

One common way to think of moral obligation is by analogy to legal obligation, and to think of both on the model of commands. We have a legal obligation to do something if we are required to do it by law, where law in turn must be issued by an appropriately authoritative person or group—a sovereign perhaps, or a duly elected legislature. If *moral* obligations are to be understood as commandments, however, there is a question about who is doing the commanding. The two most obvious candidates—God and society—are both deeply problematic.

[3] Robert M. Adams, *Finite and Infinite Goods* (New York: Oxford University Press, 1999), 250.

One problem with the Divine Command Theory is that God—more specifically, a God who gives commands—may not exist. A second is that even if God exists, God's commands to us are not easily discerned. Even if we put such metaphysical and epistemological concerns aside, however, we have reason to look elsewhere for an account of moral obligation. Whether or not we have any moral obligations does not seem to depend on the question of God's existence—it seems, for example, that we are morally obligated to refrain from killing, stealing, lying, and so on, whether God exists or not. Moreover, such obligations do not seem to be obligations *to* God, but to each other.

This last consideration may count in favor of what we might call the Social Command Theory of obligation—the view, that is, that our moral obligations come from the demands or expectations of society. However, difficulties with this view are also considerable, and may appear as insuperable as those that beset Divine Command views. For one thing, just as people may be skeptical of the existence of a commanding God, people may question whether there really is such a thing as society. Unlike the question of God's existence, the question of society's existence is not metaphysical, but it is a legitimate and serious question none the less. To be sure, we live among other people—in a neighborhood, a state, a world. But is any collection of them sufficiently organized and unified to constitute a group that can be seen to issue commands in the requisite sense? Moreover, if there is a sufficiently unified and organized group of this sort, it is far from clear that for each of us there is exactly one of them. To the contrary, we seem to be part of many different overlapping social groups—is just one of them authoritative? If so, which one, and why?[4] Furthermore, even if it be granted that there is such a thing as 'society', its commands, if such there be, are hard to discern, raising again a difficulty that runs parallel to one that afflicts Divine Command Theories. Finally, and, I think, most powerfully, there is an objection to the Social Command view that has no parallel in our assessment of Divine Command views: in so far as we admit the sense and content of the idea that society issues commands, we must acknowledge that the commands it issues are frequently mistaken. Some of the acts that societies have taken to be morally obligatory have in fact been

[4] See Andrew Oldenquist, 'Loyalties', *Journal of Philosophy*, 79 (1982), 173–93, for an interesting argument against the idea that the claims of the widest group to which we might have allegiance are necessarily also the weightiest.

morally horrific; and, some acts that we now think *are* morally obligatory, society has failed in the past, and perhaps still continues to fail, to demand.

In the light of these objections, it seems reasonable to look to some other way of understanding the concept of moral obligation, or, alternatively, to do without the concept altogether. But these other alternatives are problematic, too, as I shall now try to show. Understanding the problems behind these other alternatives will explain why I want, in the rest of the paper, to revive the Social Command Theory despite the very serious objections to it that I have already noted.

Doing Without the Concept of Obligation

Were we to give up on both Social Command and Divine Command Theories, what other accounts of moral obligation might we propose? G. E. M. Anscombe famously argued that the idea of a moral obligation made no sense if it could not be understood in terms of a command issued by an authoritative person or group.[5] Since, for reasons such as those I have given, she felt that the candidates for such an office were lacking in contemporary society, she concluded that continued talk of moral obligation was incoherent and should be eliminated. In other words, according to Anscombe, it makes no sense to say that one is morally required to do something unless one thinks that one is required *by* someone, and in particular by someone with sufficient and relevant authority. If one does not believe anyone stands in a position suitable for the occupation of that role, one should give up talking about moral requirements altogether. In that case, Anscombe thinks, one can recommend some actions and criticize others. One can talk about virtue and vice. But we should not talk of actions as being required, permitted, or forbidden, if we do not think there is any appropriate agent issuing the requirements or granting the permission.

Adopting Anscombe's suggestion would dramatically change the way we think and talk about ethics—but is there anything wrong with her proposal?

Some might be tempted to say that what is wrong is that there *are* moral obligations, and so to give up talking about them is to give up trying to

[5] G. E. M. Anscombe, 'Modern Moral Philosophy', *Philosophy*, 33 (1958).

describe and understand an aspect of moral reality. Certainly, discussions of moral obligations do not seem nonsensical or incoherent, even when one self-consciously attends to them. It would beg the question to regard this as an objection to her view, however. Her point is to challenge those of us who are not divine command theorists to make sense of our talk of obligations and requirements, and we can hardly meet the challenge by simply insisting that it does make sense, never mind how.

None the less, there are good reasons of a more practical sort for regarding Anscombe's proposal as something we should adopt only as a last resort. There are good reasons, that is, for *wanting* our moral and ethical framework to contain a distinction between the obligatory and the morally desirable, reasons that have been pointed out, for example, by Thomas Hobbes and by John Stuart Mill. Specifically, there is much to be gained—for each of us individually, as well as for society collectively, for all of humanity, and even for the whole of sentient creation—from being able to insist that people behave in certain ways and restrain themselves from behaving in certain others. The ability to insist that people deal honestly with each other and refrain from violence and theft, and the ability to insist that people come to each other's aid, if we can make our insistence effective, allow us to live in relative security, to pursue our individual goals more efficiently than would otherwise be possible, and to live in a climate of public mutual trust and respect that enhances the quality of our lives. The ability to insist on more particular forms of behavior and more particular kinds of restraint allows us to create and maintain public goods that would be impossible without coordination.

Legal systems have presumably evolved in large part as a way of fulfilling these functions, but not all the things it would be desirable to insist on are best handled through the arm of the law. Further, there are advantages to having the members of society recognize nonlegal reasons as well as legal ones for obeying some of what is properly included in the law's domain. In short, there are enormous social advantages to being able to appeal to the idea of moral obligation.

At the same time, we cannot and should not expect people to devote themselves entirely to the common good, or to constrain their own actions to a degree that would deprive them of seeking and attaining lives in which they themselves can flourish. It is in society's, or if you like, humanity's interest, that people be pressed to guide and constrain their actions in ways

that foster the common good and contribute to a climate of mutual respect, but there are also reasons—some but not all of which also have to do with promoting the common good—for wanting the scope and content of these demands to be limited.

We want people to constrain and guide their actions to some degree in order to foster the common good and to treat their fellow creatures respectfully, but we also have reason not to want people to think they must devote themselves entirely to these ends. It would be helpful therefore to have a limited category of actions that people can be expected to feel they *must* perform (and, relatedly, to have a limited category of actions that people can be expected to feel they must avoid), that they may conceive of as 'doing their share' for society or the world.[6]

People of good will and good faith will want to do their share in contributing to the world, even though this will sometimes involve some sacrifice of their own good or of their own interests. In order to raise and educate people to grow up with the concept of 'their share' and with the desire to do their share, however, we need to appeal to a distinction between what is obligatory and all else that is morally good.

Moreover, it would be helpful to appeal to such a distinction to determine how much and what kind of pressure may be put on others who may not be internally motivated to be morally good. When is it appropriate to insist that others behave in morally desirable ways? For what sorts of behavior is it reasonable to blame them? It seems intuitively inappropriate and unreasonable to blame them for failing to be as morally good as possible, but it would be very useful to be able to say of activities within certain limited ranges, that *these* are morally required, these others morally forbidden.

We have reason, then, to want a moral or ethical framework that has room for a concept of moral obligation. We have reason, that is, to find some alternative to Anscombe's proposal that we simply do ethics without that concept. But on what basis can a line between the morally obligatory and the morally good but not required be drawn, and whence would the authority of the demand to fulfill one's obligations derive?

[6] I have discussed these issues elsewhere. See esp. 'Above and Below the Line of Duty', in *Philosophical Topics*, 14 (1986), 131–48; and 'The Role of Rules', in Walter Sinnott-Armstrong and Robert Audi (eds.), *Rationality, Rules, and Ideals: Critical Essays on Bernard Gert's Moral Theory* (Lanham, Md.: Rowman and Littlefield, 2002).

Those who would reject Anscombe's proposal must either return to the option of finding an authoritative person or group who can be understood to be the source of moral demands, or dispute Anscombe's claim that the coherence of the concept of moral obligation depends on belief in the existence of such a source. We have already seen reasons for being pessimistic about the former option. In any event, the latter seems to me to be the more popular path in contemporary ethics. It involves defending the idea that the distinction between what is required and what is merely recommended can be made without reference to any commander. To say that something is morally required, in this view, is not to say that it is required *by* anyone. But then, what are we saying when we say of some act that it is required rather than merely recommended?

No-Command Theories of Obligation

To be frank, I think that much of the time we don't really know what we're saying. Often we want simply to urge someone or some group to behave in a particular way; we are not focusing on the question of whether we think the action is required or merely desirable. As I mentioned before, we speak loosely, and freely, and there is nothing wrong with that. But if we use such words as 'wrong' and 'morally required' loosely, we should not draw implications from them that would be warranted only if the words are used in a stricter, more careful way. What can the stricter, more careful use of such words as 'wrong', 'morally required', and 'morally obligatory' be?

If the distinction between what is required and what is good but not required that these words mark out is to serve the purposes I earlier mentioned, it must be associated with a distinction between the amount or kind of pressure it can be appropriate to exert in the interest of achieving conformity to certain rules or patterns of behavior. John Stuart Mill seems to have hit the nail on the head when he wrote

> We do not call anything wrong unless we mean to imply that a person ought to be punished in some way or other for doing it—if not by law, by the opinion of his fellow creatures; if not by opinion, by the reproaches of his own conscience . . . It is a part of the notion of duty in every one of its forms that a person may rightfully

be compelled to fulfil it. Duty is a thing which may be exacted from a person, as one exacts a debt. . . .[7]

Mill goes on to say

There are other things, which we wish that people should do, which we like or admire them for doing, perhaps dislike or despise them for not doing, but yet admit that they are not bound to do; it is not a case of moral obligation; we do not blame them, that is, we do not think that they are proper objects of punishment.[8]

But what can put an act into the category of the morally obligatory if its being in that category is to imply that it would be appropriate or justifiable to punish someone for failing to perform it? What can be the basis for the claim that an act is morally required if an act's being morally required is understood to license a special kind of pressure?

We have already considered briefly the suggestion that an act can be morally required only if it has been commanded by some authoritative person or group. We want now to consider whether a sense can be given to this claim that does not rely on any sort of commander. We want, in other words, to consider the possibility of a No-Command Theory of obligation.

One way the phrase 'morally required' may be used, and perhaps *is* used by some moral philosophers, identifies 'what is morally required' with 'what morality requires', which in turn can be identified with 'what consistency with moral values and ideals demands'. In the background must be some general agreement about which values and ideals are the moral ones. There are important controversies here, but in this essay I want to bracket them and assume that such agreement can be reached.[9] In other words, I want to assume that we can agree at least roughly on what counts as the realm of the moral independently of any commitment to or knowledge of the peculiar category of the morally obligatory, and that we can think of this realm as supplying us with what we can call moral reasons. Against that background, let us return to the suggestion that a moral requirement is simply a requirement *of* morality, a conclusion that follows from a commitment to moral values and ideals, or from what might be called 'the moral point of view'.

[7] John Stuart Mill, *Utilitarianism*, ch. 5 (orig. pub. 1861); Hackett edn., 1979, pp. 47–8.

[8] Ibid., 48.

[9] Moral values, presumably, include value in human life and human welfare, in treating people respectfully, in recognizing oneself as just one person among others.

Ready parallels to such a use can be found in connection with other practical concerns. Just as morality requires us to tell the truth, etiquette requires us to pass the port to the left, good spelling requires us to spell 'separate' with an 'a' (actually with two), prudence requires us to get regular dental exams. Similarly, we use the word 'must' not only in moral contexts but in non-moral contexts as well: 'You must read Hegel'; 'you must see the new Almodovar film'; 'you must visit Venice before it sinks into the sea'.

These non-moral uses of 'must' and 'require', however, have nothing to do with obligation as we ordinarily understand that term. They are rather expressions of what Kant called hypothetical imperatives. To say 'X requires Y' in these examples is to say that you should Y if you value X, or perhaps that not Y-ing is inconsistent with valuing or caring about X. The admission that X requires Y, however, only gives you a reason to Y *if* you care about X, and it may only give you a *strong* reason to Y if you care about X and about being consistent with your value of X, *very much*. When I say to someone that she must see the new Almodovar film, I do not mean to say that she has an obligation to see it. In so far as moral musts and moral requirements are understood along parallel lines, they too cannot be understood to express judgments about moral obligation.

It is tempting to express this point by saying that when we claim that X is morally required, we do not mean to say '*Morality* requires X', as we might say 'Etiquette requires Y' or that 'Good spelling requires Z'. Rather, we mean to say that *we* require X, on moral grounds or for moral reasons. But who are 'we', and on what authority are 'we' able to require anything? These questions bring us back immediately to consideration of the Social Command Theory. But there is another proposal for a No-Command Theory of obligation that I want to consider first.

If understanding moral obligations as those practical claims that are required *by morality* (or by the moral point of view) is not strong enough to capture the normative force of the concept of obligation, it may be more promising to think of moral obligations as requirements of *reason,* or, more precisely, as requirements of reason in cases in which moral considerations are decisive. (We would not consider someone morally obligated to reach correct arithmetical conclusions while balancing her check book, although such conclusions may be required by reason, too.) According to this proposal, the claim that one is morally required may be used to single out

cases in which moral reasons not only count in favour of something, they count decisively. That is, they outweigh all other reasons that might favor doing anything else. To say one is morally required to do something in this sense, would imply that, all things considered, you should do it, in a situation in which the salient reasons for this judgment are ones of a moral sort.

I suspect that I, as well as others, often do use the phrase 'morally required' this way, to refer to what I take there to be decisive moral reason to do, especially when I am considering what to do, or which moral judgments to apply to myself. However, it would be a mistake to identify this use with the concept of the morally obligatory, in so far as the latter's connection to the appropriateness of social pressure and of blame toward others is to be retained and preserved. I will offer five reasons why.

First, it can be extremely difficult to know what one has decisive reason to do, particularly when one is asking the question about another person, whose non-moral reasons one may not be in a position to assess, and when the expectations of society are indeterminate or unclear.

Second, even if we do know that another person has decisive moral reason to do something, it is not clear that this puts us in a position legitimately to stand in judgment on him or to issue blame. We would not, for example, regard it as appropriate to pressure someone to obey the rules of his club or even his religion, or to judge him for failing to do so, if we were not also members of the community in question (though we might point out to him that he was violating those rules). Why should there be no restrictions of this sort in the case of moral rules?

Third, failures to do what one has decisive reason to do in cases in which the salient reasons are *non*-moral neither do nor should result in the kind of guilt and blame appropriate to breaches of moral obligation. I have decisive reason to put down the Sudoku puzzle I am trying to solve at midnight and get to bed; I have decisive reason to set the alarm for seven, to get up, eat breakfast, and get down to work; I have decisive reason to pass up dessert; to exercise; maybe, even to read Hegel. I do none of these things, thus showing that I frequently do not do what I have decisive reason to do. I am not even close to being perfectly rational. But what of it? Such failures of rationality on my part do not seem to license blame. You would not be justified in trying to compel me to act otherwise, and I myself need not feel guilty (though I may be self-critical and displeased with myself in

other ways) for such failures of rationality. If it is appropriate—as we are insisting that it is—to blame someone for a failure to do what is morally obligatory, the fact that her failure constitutes a violation of decisive (or all things considered) reason will not explain why.[10]

A fourth reason for not identifying what one has decisive moral reason to do with what one can appropriately be punished or blamed for not doing is that in the absence of publicly expressed social expectations a person may not know what she has decisive moral reason to do. It is generally considered unfair or unjust to blame someone for failing to do what she could not reasonably be expected to know she should have done, and this case seems to be no exception.

Fifth, and finally, there is a range of cases in which my intuitive judgments about what one has decisive moral reason to do and about what is morally obligatory come apart. I suspect that many others have intuitions similar to mine in such cases. Consideration of the kind of case I have in mind may also illustrate and reinforce some of the other reasons just listed for thinking that the two categories should not be conflated.

Moral Obligations and Decisive Moral Reasons—Some Examples

These examples all concern cases in which one might say morality is in transition, in which one group of people see—let us assume correctly—that there is decisive moral reason to do something, but the bulk of society has not yet caught up with their reasoning and their insight. It is in the nature of such cases that they are culturally bound, and the examples that work best for me may not be perfectly suited to readers from other countries or even other subcultures. Still, I hope that they are suggestive enough to allow those who cannot relate easily to my specific examples to find comparable ones of their own.

A particularly clear case for my purposes concerns the question of whether to own and drive a Sports Utility Vehicle (an SUV). SUVs are

[10] As Adams writes: 'To the extent that I have done something morally wrong, I have something to feel guilty about. To the extent that I have done something irrational, I have merely something to feel silly about—and the latter is much less serious than the former', *Finite and Infinite Goods*, 238.

large cars—classified actually as small trucks—that are higher and larger than ordinary cars and typically less fuel-efficient. They are also less safe for the drivers and passengers of these cars themselves and a safety hazard to others. They are very popular in the United States. These vehicles have significant advantages for drivers who need to transport six or more people, or travel on rough terrain, but for most people who drive them such benefits are negligible or non-existent. Rather, people just like the feel of driving a big, high car—it makes them feel safer, though this is an illusion. Above all, these cars are in fashion.

Now, it seems to me, and to many of my friends, that people who have no special needs to which SUVs particularly answer should not drive them. The danger they add for other drivers and passengers, the extra damage that they cause to the roads, their unnecessary use of scarce resources, constitute decisive moral reasons to choose another car, not to be outweighed by considerations of fashion and of just liking the way driving an SUV feels. When my daughter, to whom my husband and I had promised a car, expressed a preference for getting an SUV, I said 'over my dead body' (though I did not mean to use that phrase literally). The point is, I believe that there is decisive moral reason not to buy an SUV—but I do not think it is morally obligatory.

As I mentioned, these cars are very popular. Many of my children's friends and their parents drive them. Indeed, some of my own friends and relatives drive them as well. I wish they wouldn't, and, depending on the closeness and the quality of my relationship to them, I might make a point of letting them know what I think and urge them to change. Still, I don't believe that they have a moral obligation not to buy or drive an SUV. I don't think that when they do buy them and drive them their behavior is literally immoral. It is not that I think merely that it would be inappropriate to *tell them* that their behavior is immoral—that, as it were, this is a case where it is best to keep my opinion to myself. Rather, I do not believe that it *is* immoral—they are under no obligation to choose a smaller car. Recalling Mill's distinction between what we regard as someone's duty and what 'we wish that people should do, which we like or admire them for doing, perhaps dislike or despise them for not doing, but yet admit that they are not bound to do' this case, for me, falls clearly on the latter side.

In countries where the roads are smaller, the parking is scarcer, where fuel is more expensive, and public transportation more convenient, there

may be less need to be concerned about the morality of driving an SUV. Prudential reasons may be sufficient to make the option unattractive to most. And in places where the possibility of wanting or needing to drive on rough terrain is greater, the reasons not to buy an SUV may be less decisive. But other cases in which the moral demands of society appear to be in transition may translate better to non-American contexts. To what extent do we have an obligation to recycle, to refrain from eating meat, or from purchasing eggs that are not hatched from free-range chickens? Is exclusive use of the male pronoun to refer inclusively to men and women immoral? What about the use of the word 'girls' to refer to the members of the twenty-one-year-old female field hockey team?

Some will no doubt question whether there is, as we might say, '*anything* wrong' with the patterns of behavior I have just mentioned. That is to be expected, since, as I have said, they concern issues in which social opinion is in transition. My interest, however, is especially directed toward those cases in which one is inclined to agree that there *is* something wrong with the behaviors in question, that they are morally undesirable or bad, that they should be discouraged, but yet one is hesitant to apply labels like 'morally forbidden', 'immoral', or 'wrong' to them. If one acknowledges an inclination to think that there are such cases, then it is at least plausible to consider the possibility that the category of acts we have decisive moral reason to perform is not the same as the category of the morally obligatory. But then what might the difference between these categories be? In virtue of what might an act be morally obligatory if not in virtue of there being decisive moral reason to obey it?

An answer that will be immediately suggested by the range of examples I have just been discussing is 'in virtue of the demands of society'. For a salient feature of the range of examples under consideration that may explain why even those who agree that there is decisive moral reason to conform to certain rules of behavior are reluctant to apply terms like 'morally obligatory' to these rules is that society at large has not endorsed or demanded that we conform to them. Reflection on these examples, in other words, may suggest that there is something to the Social Command Theory of moral obligation after all.

Before reconsidering that view, however, let me briefly discuss an alternative hypothesis, according to which the relevant difference between acts that are morally obligatory and those that are not has less to do with

social expectations and demands than it does with 'moral seriousness'.[11] If the claim that an act is morally obligatory is to license social pressure, guilt, and blame, we should reserve the term's application to conduct in which something of considerable moral significance is at stake. Among those who agree that there is decisive moral reason not to drive an SUV or use sexist language, disagreement about whether one is morally obligated to avoid these activities may reflect differences in the assessment of the seriousness of the matter. I am doubtful, however, that the consideration of moral seriousness will yield a satisfactory account of the distinction we are after.

There are plenty of matters of considerable moral significance that do not fall within the realm of the morally obligatory: should one use one's Christmas bonus to buy oneself a vacation or contribute it to famine relief, where it will supply a family with food for three months? Should one set aside four hours of one's busy week for workouts at the gym or use them volunteering to be a Big Sister for an inner-city teenager? There are also plenty of matters that do fall within the realm of the obligatory that are none the less of minor concern. In ordinary circumstances it is wrong, albeit trivial, to use one's employer's business stationery for personal use, or to shoplift a pack of chewing gum from a supermarket.

One might respond by suggesting that the first cases are not cases of moral obligation because the moral reasons in favor of contributing to famine relief or becoming a Big Sister are strong but not decisive. Being able to take a vacation or to maintain an exercise routine may be very important to a person, after all, and so it may be unclear where the balance of reasons lie in these cases. On what basis, however, might it become clear where the balance of reasons lie? What will determine that a strong moral reason is not just strong but decisive? At least in part, the answers to these questions seem to depend on what society expects or requires of us.

Even harder than accounting for cases where something morally serious is at stake that fall outside of the realm of obligation is the problem of dealing with cases that seem to be within the realm of obligation, but that are not morally serious. On a view that takes moral seriousness to be a condition of obligation, there can be no such thing as a trivial obligation. According to such a view, the sorts of acts I mentioned as fitting that

[11] Although he supports the view that obligations are grounded in (divine) social requirements, Adams also thinks that obligations are necessarily to be taken seriously. See *Finite and Infinite Goods*, 235.

category must either not be obligations or not be trivial after all. I do not see how to defend the claim, however, that it is not wrong to steal when the stakes in question are small (for both the thief and the victim). However, the alternative position—namely, that no case of stealing is trivial—is problematic as well. For one thing, such a view is in danger of being objectionably morally fastidious. Even in so far as it is plausible, however, it requires further explanation. Why should it be that any case of stealing is a morally serious matter? Again, the most plausible answer I can think of would refer back to social expectations and requirements. The seriousness of the matter stems from the fact that it is a violation of a social expectation or demand.

Even if it is true, therefore, that we do not regard something as morally obligatory unless we think that acting accordingly is a serious matter, there is reason to doubt that moral seriousness is a *condition* of moral obligation. If we adopt a Social Command theory, however, according to which moral obligations stem from social requirements, we would have an explanation for the strong (though not necessarily universal) connection between moral obligations and moral seriousness. Specifically, the fact that society expects certain kinds of behavior from us may give us much stronger moral reasons to adopt those behaviors than we would have in the absence of these expectations, and could make the question of whether to conform to those patterns of behavior a much more serious matter than it would otherwise be.[12]

The Social Command Theory Revisited

Robert Adams has also argued that Social Command theories of obligation have important advantages over No-Command theories. Most prominently, he has pointed out that social requirements can account for the motivational and reason-giving force of obligations in a way that no-command theories cannot. He writes

[12] There is a clear analogy here with law: the fact that a law requires us to do something typically gives us a much stronger reason to do it than we would have in the absence of the law, and can change the character of the significance of a failure to do it dramatically. In both the legal and the non-legal cases, the existence of an expressed demand or expectation can be seen as transforming what would otherwise be at best an imperfect duty into one or more perfect ones.

By contrast (with counterfactual claims about what ideal societies *would* demand), actual demands made on us in relationships that we value are undeniably real and motivationally strong. Most actual conscientiousness rests at least partly on people's sense of such demands.

The actual making of the demand is important, not only to the strength, but also to the character, of the motive. Not every good reason for doing something makes it intelligible that I should feel that I *have* to do it. This is one of the ways in which having even the best of reasons for doing something does not as such amount to having an obligation to do it. But the perception that something is demanded of me by other people, in a relationship that I value, does help make it intelligible that I should feel that I have to do it.[13]

Among the problems I listed for a theory of obligation that identified obligations with what one has decisive moral reason to do was the need to explain why failure to act rationally when moral reasons are salient should license attitudes and pressure that are uncalled for when rationality is breached in other contexts.[14] As Adams's remarks make clear, no such problem arises for theories that identify moral obligations with social requirements or commands. In those theories, a failure to discharge one's obligations constitutes a disruption of social relations, a breaking of faith or of allegiance with one's society. It is easy to understand how this can appropriately lead to guilt in those who care about being on good terms with their society and anger or criticism on the part of those who identify with the society that has been ignored or defied.

For related reasons, the Social Command Theory can explain what gives us the authority to sit in judgment on others, and to issue blame. Most of the time, we will have that authority in virtue of being members of the very society whose demands or expectations are flouted—when we try to compel someone to fulfill his obligations or blame him for failing to do so, we speak as representatives of society, or on society's behalf.

A third concern I mentioned as a problem for the view that identifies obligations with what one has decisive moral reason to do is also easily dealt with on a social command view. Specifically, I noted that if the concept of moral obligation is to be understood to be tightly connected to

[13] Adams, *Finite and Infinite Goods*, 246.

[14] It is tempting to think that an explanation can be found in the fact that failure to act on decisive moral reason is likely to harm other people, whereas failure to act on other sorts of reason is not. However, as Adams points out, not all wrong acts, for which guilt and moral blame are appropriate, do harm people. It is easy to imagine cases, for example, of plagiarism and other sorts of lies that do not.

the idea of justifiable blame, then it must be assured that in general people can be expected to know what their moral obligations are. It seems to me unreasonable to expect people always to know what they have decisive moral reason to do. Moral reasoning is hard, and separating good reasons from bad sometimes requires more sensitivity and wisdom than most of us have. Social demands and expectations, on the other hand, must be readily accessible—for unless a principle or rule is publicly expressed, and general consensus about it ascertainable, it will not satisfy the conditions of being a social demand or expectation in the first place.[15]

If these considerations speak in favor of a Social Command Theory, however, they do nothing to lessen the seriousness of the objections to that view that I mentioned earlier. If we are to understand our moral obligations as issuing from the demands of society, we need to deal with questions about what counts as a sufficiently unified and organized community to deserve the name of 'society' and about the identification of directives sufficiently articulated, forceful, and clear to deserve to be interpreted as social 'requirements' or 'demands'.[16] If, as seems likely, the answers to these questions allow that we may be members of multiple, overlapping societies, we find ourselves with a further problem—namely, that we may find ourselves subject to demands and expectations that, although separately reasonable, conflict with each other.

These questions raise difficult and important issues for a Social Command Theory of obligation, but I shall leave them for another day (if not another author). If we are in the end to accept a Social Command Theory, I see no way of avoiding the conclusion that in many cases the application of the concept will be regrettably indeterminate, due to the vagueness both of what counts as society and of what counts as a sufficiently unanimous and publicly articulated demand. At the same time, our ability to engage

[15] Adams also discusses the advantage of a Social Command Theory in ensuring that moral obligations are publicly recognized (*Finite and Infinite Goods*, 247).

[16] In this essay, I use the label 'Social Command Theory' to refer to views that take social demands and expectations to be the source of moral obligations in order to highlight the relationship such views have to Divine Command Theories of obligation. However, the core idea of such views is that obligations are a function of a social relationship, and arise out of the demands or expectations of a person or group to which one is bound by membership, gratitude, or respect. The idea that society issues guidelines in a way that can be understood specifically as commands plays no essential role in these views, and indeed seems to me somewhat forced. In a more careful and elaborate defense of the view I am proposing here, I would drop the word 'command' in favor of 'demands and expectations' or 'requirements', which is the term Adams uses.

in discussions about the cases in which society is in moral transition shows that talk of the expectations of society is not totally empty. When we ask what society expects of us we are able to reach considerable agreement on the answer, and to provide reasoned support and criticism of each other's opinions when we disagree.

Indeed, were this not so, the final, and to my mind most serious, objection to Social Command Theories could never be raised. That objection concerns the fact that society has often gotten our moral obligations wrong. Sometimes society has claimed that its members are morally obliged to do things that we now see were not morally obligatory. In other instances, society has failed to acknowledge moral obligations that its members none the less seem to have. These objections could not be meaningfully raised if we could not understand what was meant by 'society' at all, or if we had no idea what commands, or social expectations, or allegedly moral values society had.

Let us then bracket the problems having to do with the more precise specification of society and of the identification of its commands, and turn to this last objection. We may divide the objection into two parts. On the one hand, it appears that societies have sometimes demanded, in the guise of moral obligations, actions that were not in fact morally obligatory: they have demanded that people refrain, for example, from homosexual activity and masturbation, and that they refrain from sexual activity altogether outside the confines of marriage. Worse, it seems that societies have sometimes demanded that its members actively harm people who are in one way or another regarded as outsiders: Americans were taught that they were morally obligated to report runaway slaves during the era of American slavery; Germans were told that they were obliged to turn in Jews whom they knew to be hiding from the authorities. On the other hand, societies have failed to recognize as morally obligatory what many would say *are* our actual obligations: Arguably, Germans were not only *not* obliged to report hidden Jews, they *were* obliged to protect them or to help them escape. Slaveowners, it may be said, were under a moral obligation to free their slaves. And perhaps we are all obligated to do more to aid those in desperate need than any of our societies require of us.

The first part of the objection, we might say, concerns societies' tendency to issue 'false positives'—to regard certain forms of behavior as morally obligatory when in fact they are not morally obligatory and may even

be morally bad. This objection applies to what one might call a *pure* Social Command Theory, according to which the existence of a social command, conceived by society as a moral command, is both necessary *and sufficient* to establish a moral obligation on the part of that society's members. I see no reason to advocate such a view, however. The reasons for taking a Social Command Theory seriously are reasons for thinking moral obligations depend on the existence of social demands as one of its conditions. Nothing speaks in favor of insisting that no other conditions need to be met. Since the claim that someone has a moral obligation is a normative claim—since it endorses the idea that the person has a certain kind of reason to comply, and that it would be appropriate to try to compel him to comply or to punish or blame him for failing to do so—a second condition seems necessarily built into such claims. Specifically, if a demand of one's society is to give rise to a moral obligation, that demand must be one that is supported by strong moral reasons. Whether the demands meet this condition may well be largely independent of that society's own values and beliefs. This second condition will eliminate false positives—for if a social demand is not supported by good moral reasons then it will fail to give rise to a genuine moral obligation. The problem of false negatives, however, must be treated another way.

Surely, it might be argued, what moral obligations we have does not depend on the endorsement of society. Some acts, it seems clear, are morally impermissible regardless of what society says. Societies, as we know, have failed to forbid acts we all recognize as morally appalling—like genocide. According to the Social Command Theory, it appears that the members of such societies had no obligation to refrain from participating in genocide. But this seems a *reductio ad absurdum*. Any theory that fails to recognize genocide as morally intolerable has to be wrong.[17]

Before accepting this argument, it is worth examining whether, and, if so, in what kind of case, a Social Command Theory would have such an implication. Has there ever really been a society that has not regarded genocide as morally forbidden? Certainly, there are all too many cases

[17] This line of criticism constitutes one of Robert Adams's strongest objections to (human) Social Command Theories, and one of his strongest reasons for finding the move from human to Divine Command Theories of obligation compelling. According to Adams, 'moral reformers have taught us that . . . things that were morally required were not actually demanded by any community . . . In this way actual human social requirements fail to cover the whole territory of moral obligation' (*Finite and Infinite Goods*, 248).

in which societies have failed to punish genocide, in which indeed the government itself has engaged in genocide and others have cooperated without reproach. But the moral codes of these societies in recent times and even in recent centuries have all included principles that are clearly inconsistent with such behavior. In so far as the spokesmen of such societies have tried to offer moral justifications for their practices, they have all been in bad faith.

These are cases, in other words, in which the societies in question have not practiced what they preached. In such cases, Social Command Theories must specify whether it is the practice or the preaching that is weightier in determining what constitutes society's demands. Though the issues raised are more complex than I have the space to discuss here, it is at least arguable that the best Social Command Theories will take moral obligation to be grounded in society's declared moral values, that is, in the moral rules and principles that the society publicly endorses and asserts—in its schools, its newspapers, its religious institutions, and in other vehicles of cultural expression. If society does not practice what it preaches, in this view, it is what society preaches that is weightier in determining what society commands in the relevant sense. The appeal of this version of the social command view can be overlooked if the connection I have often invoked between claims of obligation and the appropriateness of compulsion and punishment is misunderstood.

It has been argued by some, for example, that it makes no sense to say that one has an obligation to do something unless there is some actual sanction that attaches to the failure to comply—that, in other words, to say that one is morally obligated to do something commits one to the thought that one must do it *or else*. Under that construal, a society that does not punish a mode of behavior cannot be said to regard restraint from that behavior as morally obligatory. But Social Command Theories are not committed to this interpretation. Rather, they may associate claims of obligation with the condition that it would be *appropriate to* put pressure on someone to conform to what is obligatory, and *appropriate to* blame him for failing to comply.

No doubt society's actual approval and disapproval, its rewards and punishments, are essential to the development of moral sensibility in its members. But, once that sensibility has been developed and moral demands appreciated and internalized, individuals are able to recognize

moral obligations and reasons to conform to them that do not depend on sanctions or any other sort of social enforcement. Once an individual has internalized society's demands, his compliance need not be motivated by fear of actual reprisal, not even by fear of reprisal by his own conscience in the form of feelings of guilt. His choosing to satisfy society's moral norms may rather be an expression of his allegiance to society, his desire to live up to its expectations, his willingness to do his share.[18]

When we discuss societies that have engaged in genocide, enslavement, and other moral abominations, we often refer to the fact that the practices in question are condemned by the society's own moral code. (Many such cases occur, for example, in societies that are predominantly Christian, some of them even in the name of Christianity, despite the central roles Christian ethics accords to the commandment against killing, the injunction to love thy neighbor, and the Golden Rule.) In so far as we regard blame and punishment for these practices to be appropriate, we rely on the thought that the participants in these practices were in a position to know better. They had the basis in their own moral training to recognize that what they were doing was wrong.

A better test of the Social Command Theory against our most recalcitrant moral intuitions might be found if we can discover or imagine a society that engages in morally horrific behavior without bad faith or hypocrisy. Has there ever been a society that did not even nominally disapprove of genocide or of the murder of innocents? If there were, and if the society committed such acts, wouldn't it still be true that what they did was wrong? I am not enough of a historian to know the answer to the first question, but I confess that when I hear descriptions of the morally horrible practices of sufficiently distant cultures, I do not find it natural to think of such events in terms of the vocabulary of right and wrong, or of moral obligation.

Recently, for example, I was told that the reason Oxford has no buildings more than a thousand years old is that a Danish king had it razed to the ground to avenge himself against Ethelred the Unready, who had ordered all the Danes in England slaughtered, in the course of which the Danish king had lost his son. This was horrible behavior, surely, on the parts of both Ethelred and the Danish King, but it is hard to see the point of judging whether either of them violated any moral obligations in acting as they did.

[18] Similar comments apply to versions of the Divine Command Theory.

Such language seems not only too weak but irrelevant to the interest such historical events hold for us.[19]

Contemplation of societies that lack commands against behavior we would regard as morally outrageous, then, does not generate in me any intuitions that count against the Social Command Theory. Such societies are apt to be very distant from ours, and the reasons we might have for using the language of moral obligation to describe them seem very weak. Perhaps more important, though, is the fact that the concept of moral obligation, when it is used strictly to mark off the category of acts upon which pressure to conform is legitimately put, is a limited and specialized concept, not to be confused or identified with other forms of moral assessment.

If acceptance of a Social Command Theory of moral obligation prevents one from being entitled to say that the members of an extremely historically distant society violated their moral obligations when they acted in morally horrific ways, it does not prevent one from saying all sorts of other things in criticism of that society's practices. It does not prevent one, for example, from being entitled to say that the practices were morally abhorrent; that the victims of these practices were abominably treated; or, that to live in such a society would have been in certain respects very bad. Nor does it prevent one from being entitled to say that there is decisive moral reason to resist such practices—at least for us, but also possibly for them.

If one wants to reserve the concept of moral obligation for those acts that it would be appropriate to compel someone to do, and to punish someone for not doing, then it seems to me that a Social Command Theory offers the best account of that category. But then we should also admit that one may have decisive moral reason to do something even though it is not morally obligatory—and, in many contexts, that is the question we ought really be trying to answer.

Deconstructing the Concept of Moral Obligation

It might be objected that my defense of the Social Command Theory of moral obligation has come at the cost of depriving the term 'moral

[19] These remarks echo Bernard Williams's position about the relativism of distance in *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press, 1985), 162–5.

obligation' of much of its interest and power, for I have insisted on reserving the category of the morally obligatory for a rather specialized use. Once one realizes how specific a function I understand the concept to play in our moral framework, one might wonder why one should care what our moral obligations are, or why one should care about formulating a proper account of the category of the morally obligatory.

I am not wholly unsympathetic to this response, though I prefer not to see it as an objection. Instead of characterizing what I have been doing in this paper as defending an account of moral obligation, it would perhaps have been just as well to describe my project as that of sorting out the strands of our present use of the term 'moral obligation', and showing how the connections among these strands cannot always be guaranteed.

On the one hand, we sometimes identify the morally obligatory, as Mill does, with that which it would be appropriate to compel someone to do, or to punish or blame him for not doing. When we say of someone 'He is under a moral obligation to do X', we implicitly *do* blame him, or at least license blame, if he fails to comply. On the other hand, perhaps we sometimes use 'morally obligatory' when we think a person has decisive moral reason to do something—when we think, that is, that there are strong moral reasons for him to do it which outweigh whatever non-moral reasons he might have in favor of doing something else.

Of course, these two uses will often overlap, for if an act is such that it would be appropriate to compel someone to do it, there must be at least a strong moral reason for doing it, and the chances that the reasons are so strong as to outweigh any competing non-moral reasons are good. Conversely, if there is decisive moral reason to perform an act, there may well be reason for wanting to put pressure on people to perform it, and if there is such reason, and society lets it be known that there is, this may make it appropriate to compel someone to perform it or to blame him if he does not.

Though these two strands of our use of the term 'moral obligation' will tend to overlap, however, I have argued in this paper that they may none the less come apart. In so far as we want to restrict the term to its first use, I have further argued, we should accept a Social Command Theory of moral obligation.

Do I want also to argue that we *should* restrict ourselves to the first use? Even if I did, it would be futile for me, reticent philosopher that I am,

to try to get people to change the way they talk. I do think that there is a point to having a category that conforms to the first use, however, and that a failure to recognize the difference between this category and the category of 'what one has decisive moral reason to do' has some unfortunate effects.

One unfortunate effect has to do with the way we judge ourselves. Specifically, it seems to me that people who want to be morally decent but do not want to sacrifice more than they have to for this end sometimes ask themselves whether they are under a moral obligation to extend themselves in certain ways or restrain themselves in others. In asking this, they may be implicitly appealing to a standard appropriate to the first category—the standard according to which an act is morally obligatory only if it would be appropriate to compel someone to do it or to blame someone or punish him if he does not. But we have seen that even if something escapes being morally obligatory by this standard a person may have decisive moral reason to do it none the less. In other words, it may be true that, all things considered, the person should do it, and for moral reasons. It would be better in such cases if people asked themselves directly what if anything they had decisive moral reason to do, rather than relying too heavily and taking too seriously the concept of moral obligation.

When thinking and talking about what others in our society should do, our failure to distinguish the two strands of our use of 'moral obligation' has two other unfortunate effects. First, we are apt too readily to move from a conviction that people have decisive moral reasons to act in certain ways to a willingness to blame them for failing to act as we think they should. We are apt, in other words, to be overly moralistic and judgmental. Second, by allowing ourselves to move directly from judgments about what individuals have decisive moral reason to do to judgments about what it would be appropriate to blame them for not doing, we may locate the source of our moral dissatisfaction in the wrong place, and thus fail to see what we ourselves have moral reason to do.

Specifically, we may fail to see the situation as one that calls not for private moral judgment but for public moral action. If we believe that people have decisive moral reason to act in ways that society none the less does not demand of them, it may be that what is needed is that we work toward bringing it about that society does demand it. By writing editorials, campaigning for social change, raising public awareness, we can raise the

moral bar—that is, we can help to bring it about that behavior that is currently not morally obligatory becomes so.

This is not always desirable—the category of the morally obligatory is meant to balance society's interest in enforcing conformity to certain rules against an interest in protecting people from too many restrictions in how they choose to live their lives and pursue their goals. The moral bar may thus be set either too high or too low. Still, by recognizing the difference between what our moral obligations are and what there is moral reason to want them to be, we may have a better chance of evolving into a society in which these two categories more closely coincide.

11

Adams on the Nature of Obligation

JEFFREY STOUT

In a recent history of the analytic philosophical tradition, Scott Soames heralds the success of the tradition in 'understanding, and separating one from another, the fundamental methodological notions of logical consequence, logical truth, necessary truth, and apriori truth'. 'It is a measure of the importance of these achievements', Soames adds, 'that they have reverberated across all areas of philosophy in the analytic tradition.'[1] Robert Merrihew Adams's most important work in ethics, *Finite and Infinite Goods*, is an excellent example of the effects of this reverberation.[2] Even at those points where Bob Adams introduces assumptions I find dubious or where it seems to me that further distinctions, or somewhat different distinctions, need to be drawn, I have profited enormously from the clarity he has brought to the subject.

What is more, I think Adams is largely right about many topics in ethical theory. He is right, specifically, in holding: that rightness, wrongness, obligation, and permission cannot be understood without making reference to goodness; that one can acknowledge this conceptual dependence without abandoning commitment to a liberal democratic conception of politics; that the kind of goodness one ought to place at the center of one's reflections on how to live is excellence; that while the badness of some things is a matter of an absence of goodness that ought to be present in them, the

[1] Scott Soames, *Philosophical Analysis in the Twentieth Century*, i (Princeton, NJ: Princeton University Press, 2003), p. xi.

[2] Robert M. Adams, *Finite and Infinite Goods: A Framework for Ethics* (New York and Oxford: Oxford University Press, 1999), 273, 212. It is, I believe, the most rigorously systematic and profound theistic account of ethics to appear in the modern period.

badness of some persons, acts, or attitudes is a matter of their opposition to or hatred of something good; that a finite thing has sacred value insofar as its violation or destruction would be horrible and that genuine obligations are best understood as requirements that arise in the context of social relationships that have not been rendered defective in certain ways. The points of agreement just enumerated are so extensive and pertain to issues of such great structural importance in ethical theory that Adams and I can be described as offering divergent interpretations of a single set of inferential relationships. When discussing ethics, he and I appear to have roughly the same basic inferential structure in mind. While I am disposed to interpret that structure in what might be described as a metaphysically minimalist way, Adams is disposed to construe that structure in metaphysically ample terms. He sees my interpretation as too parsimonious to make full sense of the structure we both have in mind. I see his interpretation as too encumbered by extravagant metaphysical assumptions to be plausible.

Adams is, of course, widely credited with having pumped life into two types of ethical theory recently rumored to be dead. A generation ago few philosophers defended either theistic or Platonistic ethical theories, and most philosophers thought that a small collection of arguments from Plato, David Hume, and G. E. Moore had essentially eliminated divine command theories of ethics from serious consideration. The philosophical landscape now looks much different, thanks to Adams's work. *Finite and Infinite Goods* is simultaneously Platonistic and heavily indebted to the broad tradition of Jewish and Christian theism. The book's central Platonistic claim, which Adams shares with Iris Murdoch, is (1) that the goodness of finite things must be understood in terms of a single transcendent Good.[3] The two most important claims that Adams takes over from Judeo-Christian theism are ones that Murdoch would reject as the projections of an ego in search of consolation, namely: (2) that the transcendent Good is a loving God and (3) that moral obligations must ultimately be identified with the commands of God thus understood. In this paper I will be focusing mainly on (3), but all three claims will be at issue. One implication of my argument will be that Adams's Platonism does not sit as easily with biblical theism as he thinks it does.

[3] Iris Murdoch, *The Sovereignty of Good* (New York: Schocken Books, 1971).

The present paper continues a dialogue on these matters that Adams and I have been engaged in for many years. In 1973, he had set out the original version of his divine command theory of ethical wrongness as a set of semantic theses.[4] In the first article I published in a professional journal, in 1978, I argued that Adams's theory could not be assessed without making some assumptions in semantics.[5] If we take as the starting-point for our semantics the admittedly vague notion that meaning is use, and then try to specify what sorts of use might plausibly be thought to have an effect on an expression's meaning, inferential role will be a leading contender. For the purposes of argument, I assumed that the meaning of an ethical expression in a given language is in part a function of the roles it plays in the inferential practices of a community that uses that language. I then took Adams's claim that 'ethically wrong' means *contrary to the commands of a loving God* as a claim about the inferential role of an expression within certain theistic ethical communities. Taken in this way, however, the claim turns out to be true but neither surprising nor consequential. The communities in question do routinely treat the inference from *x is contrary to the commands of a loving God* to *x is ethically wrong* as materially sound. If meaning is partly a matter of inferential role, then this inference must contribute to the meaning of the phrase 'ethically wrong' as it is employed within such a community. It does not follow, however, that we ought to join members of such a community in treating the material inference as sound. Nor does it follow that they are correct in treating it as sound. Neither, for that matter, does it follow that Adams was right in holding that his own concept of ethical wrongness would 'break down' in the unlikely event that God—assuming that there is one—commanded cruelty for its own sake. So if Adams wanted to vindicate his metaethical claims as both true and consequential, he needed, at a minimum, to say what semantic theory he had in mind.

Meanwhile, Adams had been reading groundbreaking works by Hilary Putnam and Saul Kripke, and decided under their influence to recast his theory of ethical wrongness not as a set of theses about the *meaning* of

[4] Adams, 'A Modified Divine Command Theory of Ethical Wrongness', in Gene Outka and John P. Reeder, Jr. (eds.), *Religion and Morality: A Collection of Essays* (Garden City, NJ: Doubleday Anchor, 1973), 318–47.

[5] Jeffrey Stout, 'Metaethics and the Death of Meaning: Adams' Tantalizing Closing', *Journal of Religious Ethics*, 6 (1978), 1–18.

'ethically wrong', but rather as a set of theses about the *nature* of ethical wrongness.[6] The theories of goodness and moral obligation defended in *Finite and Infinite Goods* are similarly cast as theses concerning the natures of these things. An acceptable semantics of ethical expressions, Adams argues, tells us what they mean, and will at least have to be sensitive to how they are used. The nature of goodness is not, however, reducible to what 'good' means; the nature of moral obligation is not simply a matter of how the phrase 'morally obligated' is used in this or that community. The nature of goodness and the nature of moral obligation are, finally, metaphysical topics, not matters of semantics. A study of the use of ethical expressions can provide information about the 'role' that the nature of moral obligation plays in ethics. Different ethical theories propose different 'candidates' for playing that role.[7] In this sense, semantics places constraints on an acceptable metaphysics of morals. But in inquiring into a topic like the nature of moral obligation, we are not inquiring into a matter of usage. We are inquiring into *the essence of something to which an ethical expression refers*. Study of the use of that expression can point us in the direction of that referent, but the nature of the referent cannot be discovered in the linguistic behavior of those who use the expression.

Adams invites us to consider the distinction between questions of the nature of things and questions of meaning in the context of scientific inquiry. He summarizes Putnam's treatment of the example of inquiry into the nature of water as follows:

> It is the nature of water to be H_2O, it is claimed; and the property of being water is, necessarily, identical with the property of being H_2O. But the word 'water' does not *mean* H_2O. What I must know, at least implicitly, about water in order to understand the sense of the word 'water', and so to be a competent user of the word, is that if there is a single chemical nature shared by most of the stuff that I and other English-speakers have been calling 'water', then, of necessity, all and only stuff of that nature is water. The causal relations between concrete samples of water, on the one hand, and users and uses of the word 'water', on the other hand, serve to 'fix the reference' of the word—that is, to determine which stuff

[6] Adams, 'Divine Command Metaethics Modified Again', *Journal of Religious Ethics*, 7 (1979), 66–79. The relevant works of Putnam and Kripke include: Hilary Putnam, *Mind, Language, and Reality: Philosophical Papers*, vol. ii (Cambridge: Cambridge University Press, 1975) and Saul Kripke, *Naming and Necessity* (Cambridge, Mass.: Harvard University Press, 1980).

[7] *Finite and Infinite Goods*, 16.

the word names. But the nature of water is to be discovered in the water and not in our concepts.[8]

Adams argues that we 'use ethical terms in an analogous way, which enables us to distinguish between the semantics of ethical discourse and what we may call the metaphysical part of ethical theory'.[9]

What water or moral obligation *is* and what the expressions 'water' and 'moral obligation' *mean* should indeed be distinguished. And it is true, as Adams says, that people used the term 'water' (and its equivalents in other languages) for a long time before discovering what water is made of. I am not persuaded, however, that the question of what water is made of and the question of what water is—the question of constitution and the question of identity—are the same question. Nor am I persuaded that inquiry into the nature of moral obligation is best understood as a metaphysical discipline analogous to inquiry, in the physical sciences, into such questions as what water is made of. Indeed, I am not persuaded that what Adams calls the 'metaphysical part of ethical theory' is a sufficiently disciplined activity to entitle those who practice it to view their conclusions as findings. Most important, it is not clear to me that engaging in this activity actually gives one the kind of knowledge or understanding that would help one live well if one had it. I will come back to these concerns toward the end of the paper, but, in the meantime, I want to focus on what Adams says about the nature of moral obligation, given his assumptions about the metaphysics of morals and its resemblance to inquiry in the physical sciences.

Adams holds that obligation is most sensibly construed in social terms. Obligations are requirements that arise within relationships among persons. The trouble is that a complete theory of moral obligation must be able to sustain a distinction between those requirements that are genuinely binding on someone and those that are not, and it is not clear, according to Adams, how this distinction can be explicated in a satisfactory way while restricting oneself entirely to relationships among human persons. The relationship between master and slave no doubt generates requirements in this sense: masters *require* slaves to obey their orders. Slaves must do the work they are told to do, defer to the judgment of their masters, and even submit to the

[8] *Finite and Infinite Goods*, 15; italics in original. [9] Ibid., 16.

sexual advances of their masters if they are to satisfy the requirements of the relationship in which they find themselves, the requirements that have been imposed on them. Above all, they must not rebel; this, too, is required of them. The master is the one who determines what is required within the terms of the relationship. But, surely these requirements, though generated by a social relationship, are not morally binding. A group of slaves that rose up in defiance of their masters, violated the explicitly stated requirements of the master–slave relationship, and managed to free themselves—by just and proportionate means—would not have violated their actual moral obligations. So being morally obligated to do something cannot simply be identified with being required to do something by others to whom one is socially related.

One could argue, in a way that is consistent with Adams's position on the conceptual dependence of obligation on the good, that what is lacking in this bare-bones theory of moral obligation is an account of the difference between social relationships that are good and social relationships that are defective in some important respect. The bindingness of genuine social requirements—of actual obligations—could then be explained as a function of the goodness of the underlying social relationships. In so far as the underlying social relationship is defective, the requirements belonging to that relationship lose their normative authority over at least some of the persons involved, with the result that requirements arising in severely defective relationships might not succeed in obligating a disadvantaged party at all, from a moral point of view. It would be interesting to see how far this thought could be pursued without using theistic assumptions to reinforce the distinction between good and defective human relationships. One way of pursuing the thought would be in terms of what is good *for* the parties involved in a relationship, quite generally conceived. Another way would be to focus more specifically on the difference between relationships that involve an arbitrary exercise of power by one party over another and relationships that do not.

Adams does not explore such possibilities. His reason is that he has already proposed a theory of excellence according to which the excellence of any finite thing is identified with the resemblance of that thing to God, and this theory provides him with resources that can be exploited in a theory of obligation. Adams's theory of excellence is, from my point of view, pretty obscure. In the first place, it is not clear what Adams takes resemblance

to be. Nor is it clear how resemblance to a transcendent God can *explain to us* what the goodness of a finite thing is, in the sense of rendering such goodness less mysterious, given that divine transcendence puts one pole of the posited resemblance-relation essentially beyond the realm in which relations of resemblance *among finite things* obtain.[10] An ethical theory that first introduces a divinity by way of a sharp *contrast* between it and finite things, and then looks to the relation of *resemblance* between that divinity and finite things for an explanation of the goodness of finite things, needs some way of showing that the relation of resemblance being posited is a determinate one. Otherwise, the goodness of finite things could itself turn out to be indeterminate—an outcome Adams says he wishes to avoid.

Adams's metaphysical explanations are supposed to swing free of the question of what we finite beings can know about the transcendent divinity posited by the theory. My point, however, is that for Adams's theory to have the main advantage he claims for it, we need to be in a position to know that the resemblance between God and finite things is determinate. I am prepared to grant that *if* God exists, a resemblance between God and finite things *might* explain the goodness of finite things. But notice: if resemblance turns out to be a perspectival affair, then Adams's explanation does not make the goodness of finite things determinate; and even if resemblance turns out to be a determinate matter in most cases, it is not obvious what it comes to when the objects that are being said to resemble one another stand on opposite sides of an ontological gulf as wide as any contemplated in metaphysics. So it would seem that Adams has more explaining to do.

Assuming, however, that the excellence of finite things is to be understood in terms of resemblance to God, Adams could try to distinguish good from defective social relationships among human persons—and thus to distinguish genuine moral obligations from non-binding social requirements—by claiming that the good social relationships are those that resemble God. This is not in fact how Adams chooses to argue, perhaps because he fears that doing so would still leave moral obligation insufficiently determinate.

[10] I have benefited from reading an unpublished paper by David Decosimo that expresses doubts about the role played in Adams's theory of excellence by the notion of resemblance to God.

Adams's theory of excellence posits a God who is not only a supremely fitting object of our love, but a loving, personal being. The transcendent Good is, as Adams puts it, a lover. To arrive at Adams's theory of moral obligation, one need only add two claims: first, that this divine lover issues commands and, second, that to be morally obligated to do something is (the same thing as) to have been commanded to do it by this divine lover. Moral obligation is a matter of what is commanded by a loving God.

The significance of Adams's reference to a *loving* God can be seen in relation to the familiar dilemma posed by Socrates in Plato's *Euthyphro*. Having fallen into the trap Socrates has set for him, Euthyphro brags that he knows the essence of piety. He first offers examples of pious actions, but Socrates argues that no such examples can succeed in capturing the essence of piety. When Euthyphro says that the essence of being pious is being loved by the gods, Socrates eventually asks whether that which is pious is loved by the gods because it is pious, or is pious because it is loved by the gods. Socrates proceeds to argue against the latter possibility: the idea that what is pious is pious *because* it is loved by the gods. If Socrates is right, one cannot reduce being pious to being loved by the gods. The two properties are not identical, even if they are instantiated in exactly the same persons and acts. Some philosophers inspired by Socrates's argument for this conclusion have taken it to apply quite broadly to a wide range of reductive accounts of normative properties and statuses. Echoing Socrates, they ask the divine command theorist of moral obligation whether that which is morally obligatory is commanded by God because it is morally obligatory, or is morally obligatory because it is commanded by God. They then construct arguments, analogous to the ones Socrates offers, against the notion that an act's being morally obligatory can be reduced to its being commanded by God.

The idea is that whatever normative property or status we are setting out to account for, what that property or status is cannot simply be a matter of how God responds to persons or acts possessing it, because this will still leave us without an answer to the question of how or why God responds in this way to persons or acts that have the property or status in question. Either God has reasons for responding, which constrain him to respond as he does, or not. If he does have such reasons, they will have to enter into the explanation of the property or status being explained, in which case the proposed reduction is not the whole story, and the essence of the property

or status remains elusive. If, however, God has no reason for responding in this way—in this case, to obligatory acts by commanding them—then the normative property or status in question becomes disturbingly arbitrary unless something more is said about God's character. If nothing about God's reasons or character constrains God to respond negatively to extremely cruel and hateful acts, then God could, by fiat, command them. In that case, on the assumption that the reductively theological account of moral obligation is true: if God commanded such acts they *would be* obligatory. But, taken in this direction, the proposed reductive account becomes extremely costly, if not unacceptably counterintuitive, even in the eyes of many theists.

Adams's 'modified' divine command theory addresses this difficulty, first, by claiming to account only for obligation (and not, for example, goodness) in terms of God's commands and, second, by referring specifically to the commands of a *loving* God. The theory claims that to be morally obligated to do something *is* to have been commanded by a loving God to do it. On the assumption that a loving God has not and will not command hateful or cruel acts, the theory does not entail that any such act is in fact morally obligatory. So the modified theory does not entail the most worrisome consequences of its unmodified predecessor.

Adams's theory presupposes that a loving God exists and that such a God has in fact commanded all of the acts that most human societies have considered morally obligatory. The acceptability of the theory depends on the acceptability of its presuppositions. For an account of obligation to be vindicated, it needs to turn out to be part of the best available account of various other things on which its presuppositions have a bearing, in this case including the nature and existence of God and the existence of horrendous evils that do not result from human fault or wrongdoing. This could be called the *systematicity constraint* on accounts of obligation, and it is a constraint that Adams seems to accept. The various parts of his ethical framework are meant to cohere with each other and, ultimately, with the best philosophical theology. That, I take it, is why Adams says, early in *Finite and Infinite Goods*, that 'The book's most important omission is that it does not address the problem of evil'.[11] It is probably also why he spends a few pages trying to show that an updated

[11] *Finite and Infinite Goods*, 7.

version of Anselm's ontological argument for God's existence at least holds promise,[12] given what can be said in support of its most controversial premises.

The systematicity constraint implies that the view of God presupposed in Adams's account of obligation needs, at a minimum, to be plausible, all things considered. Now, suppose that God's love is strongly preferential and impulsive, as it is often represented as being in the Bible.[13] What, then, is to keep the loving God from commanding the protection of those whom God loves but the murder or torture of various others? The only way for a divine command theorist of moral obligation to rule out the possibility that a loving God will command the commission of cruel acts would be to say that God's love is universalistic. And this appears to be Adams's view. But if God loves in *that* way, how are we to interpret the horrendous evils we know there to be, which fall very heavily and unequally on particular human beings?

Adams might want to respond to this question by arguing in the following way. While we must not presume to know why God allows horrendous evils to befall some human beings, including some children who have apparently done nothing to deserve such experiences, God's perfect love and omnipotence combine to make God both willing and able to defeat whatever evils there have ever been and to restore to the victims of such evils a justified sense that their lives have been worth living.[14] This is a powerful and interesting response to the problem of evil, albeit one whose strengths and weaknesses can hardly be assessed here. The trouble with offering it in the present context is that it appears to open up the possibility that a loving God could command acts like murder and genocide. If a loving God is willing to *allow* such acts to occur, presumably *because* God is willing and able in the long run to defeat such horrors and heal those victimized by them, then how can we be certain that such a God does not *command* such acts, while preparing and promising the same remedy? If God's willingness to allow horrendous evils to befall some human beings does not count against the notion that God loves all human beings equally, because of God's power to defeat even horrendous evils, then why should

[12] Ibid., 42–5.

[13] I have benefited from reading an unpublished paper on this possibility by Meir Soloveichik.

[14] See Marilyn McCord Adams, *Horrendous Evils and the Goodness of God* (Ithaca, NY: Cornell University Press, 1999).

God's willingness to command the commission of such evils be assumed to be out of the question?

The connection in Adams's reasoning between his understanding of God's love and the claim that God will not command cruel acts comes out most clearly in a chapter entitled 'Abraham's Dilemma', where Adams argues that the actual God did not in fact command Abraham to kill Isaac. Adams's intuitions of what divine perfection must consist in—intuitions not shared by all monotheistic traditions or even by non-liberal Christians—play a heavy role in the argument. I do not see how his liberal revisionism can stop at dissent from the standard reading of one biblical story. It must extend to some of the other things God commands in the Old Testament, as well as to the punishments mandated for those guilty of abominations in Leviticus and for those condemned to eternal hellfire in the New Testament.[15]

Let us focus for a moment on the genocidal acts God commands the Israelites to commit against the residents of Canaan in the Exodus story. Genocide is a horrendous evil if ever there was one. Presumably, Adams holds that the actual God commanded no such thing. But that still leaves him with the following problem. In his view, God, the infinite Good, loves universalistically, but is also the creator of a world in which horrendous evils befall human beings, often through no fault of their own. God, being God, is capable of preventing those horrors, but does not prevent them. How could that be? Presumably, it is because God has a plan for the salvation and healing of everyone. But there is more: the same God chooses not to prevent millions of sincere believers from taking the Bible as God's word and thus from taking God, on the basis of that alleged revelation, to be the exact opposite of a horror-defeater—namely, a horror-commander. This, too, must be part of God's plan for human beings. On what grounds, then, are we to conclude that the commands of the actual God will not mandate at least some acts of cruelty toward human beings? If God not only permits horrors to befall innocent infants, but also *permits millions of sincere believers to accept as divine revelation a book according to which God has sometimes commanded genocide,* then what can it mean for God to be described as loving? If these choices on God's part would not qualify as cruel, what would?

[15] In the discussion of an earlier draft of this paper at the 2005 Yale symposium in Adams's honor, he granted that he does indeed reject what all of these biblical passages appear to represent God as commanding or as planning to do.

Suppose, however, that God has not commanded and will never command acts of cruelty. This would rule out that genocide, according to the theory, could turn out to be obligatory. But it would not necessarily rule out that genocide could turn out to be morally indifferent. For that consequence to be ruled out, the loving God posited by the theory needs to command that genocide *not* be committed—and so on for other acts that seem, on reflection, to be obviously impermissible. If the commands of the actual God do not include the prohibition of, say, murder, then, according to Adams's theory, we would not be morally obligated to refrain from murder. It does not suffice, then, to exclude the possibility that God will issue cruel commands. All of the moral obligations there are need to be brought under the scope of the theory.

What reason do we have for supposing that God has in fact issued commands against all of the many acts we reasonably take to be morally impermissible? In answering this question, Adams will probably need to fall back on a priori considerations concerning perfection. God, he holds, is nothing if not perfect. A perfect being would not command cruelty for its own sake. A perfect being who is both a lover and a commander would issue commands to rule out each and every morally impermissible act, including, of course, the horrendously evil ones. Is it obvious, however, that a perfect being would need or wish to issue any commands at all? A Platonist who had jettisoned the stories of Hesiod without replacing them by the stories of the Old and New Testaments might well conceive of the divine simply as the perfect One and not as an issuer of commands at all. Why should theists not favor such a conception of divinity?

Adams appears to lean heavily on his own intuitions about the nature of perfection when arguing that God could not have commanded Abraham to kill Isaac. I have argued that his revisionist approach to the Bible has quite sweeping implications, that if the binding of Isaac is to be excluded from the canon, so must many other biblical stories be excluded for the same reason. But the more rigorously Adams pursues the implications of his scriptural revisionism, the harder it will be for him to avoid the conclusion that a *perfect* being would not resemble the Bible's portrait of God much at all. Adams has not said enough to earn entitlement to his implied identification of the transcendent Good, as conceived in terms of

Platonistic perfection, with the God of historical Judaism and Christianity. The God of his ethical theory is the transcendent Good, but also an equal-opportunity lover who commands all actually obligatory acts and only such acts as a perfect being would command. The key notion turns out to be that of perfection—not as the Bible portrays it, nor as ancient Israel understood it, nor as the Christian church has understood it, but as a contemporary liberal Protestant philosopher understands it. It is a contemporary liberal's intuitions about perfection that determine the implications of the theory for disputed cases.

As for the theory's emphasis on divine commands, Adams seems to be constructing his theory by making the smallest modification of the original divine command theory that he can while still removing the source of some counter-intuitive consequences that plagued the original version. I can see why the unwanted implications counted against the unmodified theory. But I have trouble seeing why anyone not already committed to a strongly voluntarist strand of Jewish, Christian, or Muslim theism would suppose that we can arrive at the truth about what moral obligation essentially is by making a series of relatively small modifications in the divine command theory. Adams must be assuming that the alternatives to divine command theory leave the nature of moral obligation inadequately explained regardless of how they might be modified.

Non-theological versions of the social theory of obligation, according to Adams, leave us unable to portray the distinction between morally binding obligations and social requirements of other sorts as objective or determinate.[16] I have already given reasons for thinking that a social theory of obligation might well be able to make sense of this distinction by claiming that the normative grip of a given social requirement on a disadvantaged party depends on the absence of certain sorts of defects in the underlying social relationship. Whether one develops this suggestion theologically or non-theologically, divine commands need not play an essential role in the theory, provided that the presence or absence of the relevant sorts of defects in the underlying social relationships is itself a determinate matter. If we applied Adams's conception of the excellence of finite things directly to social relationships, and granted Adams's assumption that resemblance to God is a determinate relation, we would already have a way of construing

[16] *Finite and Infinite Goods*, chs. 10–11.

the difference between defective and non-defective social relationships as a reasonably determinate matter, without making reference to divine commands. So, even on Adams's assumptions, there seems to be little reason for insisting on the importance of divine commands in an account of what moral obligation is. Moreover, I do not see any reason for thinking that a non-theological version of this approach would be likely to leave this difference *unacceptably* indeterminate. If the relevant sort of relational defect turns out to be somewhat vague, this would just mean that the line between binding and non-binding obligations is somewhat fuzzy. But why suppose that it is not fuzzy?[17]

Historically, the main theological motivation for divine command theories of moral obligation has been the conviction that there is something religiously inappropriate, as well as metaphysically inaccurate, about attributing obligations to God. Critics of divine command theories of obligation have long argued that such theories leave one unable to give content to the traditional practice of praising God as just or righteous. In *Finite and Infinite Goods,* Adams goes beyond his previous work on this topic by granting that justice is indeed one excellence to be attributed to God:

> It clearly matters to the persuasive power of God's character, as a source of moral requirement, that the divine will is just. Here, if my theory of obligation is not to be circular, I must be using a 'thin theory' of justice, so to speak, which does not presuppose moral obligation as such. Without going beyond such a thin theory I can say, for example, that God judges in accordance with the facts, and cares about each person's interests in a way that is good . . .[18]

Adams claims that a divine command theorist of moral obligation can consistently praise God for being just, provided that 'just' is not in this context treated as equivalent to 'dutiful' or 'law-abiding'.

> The justice for which God is praised belongs in the first instance to the ethics of excellence or virtue rather than to that of obligation. It chiefly involves responding well to the various claims and interests involved in a situation . . . Responding well is an excellence, and God is praised as the supreme and definitive standard of it, as of excellence in general. It does not essentially involve being under obligation, and can therefore belong to God even if God is not subject to obligation in the

[17] A related form of fuzziness pertains to the vagueness of ethical concepts, a topic I have treated at length in 'A House Founded on the Sea: Is Democracy a Dictatorship of Relativism?', *Common Knowledge,* 13 (2007), 385–403.

[18] *Finite and Infinite Goods*, 254.

same sense we are. God's justice, so understood, grounds obligation, rather than being grounded in it.[19]

By adding this feature to his theory, Adams is able to respond to the charge that he is unable to make sense of the activity of praising God's justice or righteousness.[20] Adams does not, however, make clear how he wishes to interpret the biblical portrait of God as someone who makes promises and enters into covenants. In what sense does God make a promise to do something if God does not acquire an obligation to do that thing? What could it mean for God to enter a covenant without acquiring obligations?[21] To make a promise or to enter a covenant *is* to place oneself under obligations of a certain kind. So if Adams wants to retain a divine command theory of obligation, he will either have to treat the biblical themes of promise and covenant metaphorically or drop the biblical conception of God as a maker of promises and as a partner in covenant with Israel. Here, too, Adams appears to be putting himself at odds with historical Judaism and Christianity, but in this case the source of the tension is not Platonism, but rather the voluntarist strand of biblical religion itself. God cannot be both perfectly unconstrained and bound by promises and covenants. The primary theological reason for favoring a divine command theory is to conserve a conception of God's transcendent freedom, but it seems hard, in the end, to save that conception while also saving other features of a recognizably biblical portrait of God.[22]

An inquiry like Adams's begins in puzzlement about what some normative property or status is. His attempt to reduce the puzzlement leads to relatively unconstrained speculation about the nature and existence of God, and this speculation leads in turn to puzzlement of an entirely different order: how a loving God could create a world rife with undeserved suffering and horror; what commandments God has actually issued; what aspects of

[19] *Finite and Infinite Goods*, 254–5.

[20] I am not saying that Adams's critics will necessarily be satisfied with this response. See, for example, Alasdair MacIntyre, 'Which God Ought We Obey?', *Faith and Philosophy*, 3 (1986), 359–71.

[21] See Nicholas Wolterstorff, *Divine Discourse: Philosophical Reflections on the Claim that God Speaks* (Cambridge: Cambridge University Press, 1995), 95–113.

[22] I do not mean to imply that the theme of God's transcendent freedom cannot coherently be combined with the biblical portrait of God as a maker of promises and as one who enters covenants. Karl Barth, in my view, does combine these elements coherently. The sticking point in Adams's case is the divine command theory's claim that God does not have any obligations, including ones that God freely takes on.

God's character should be thought to explain God's commanding or loving of this or that; what resemblance is; how resemblance to a transcendent God could be thought to explain the excellence of finite things; what it could mean to say that God is just; and what it could mean to say that God makes promises and enters into covenants. I am not inclined to count a replacement of the original puzzlement by the latter perplexities as an advance.

Guided by the example of Socrates, I find myself thinking that it would be rash to commit myself to a set of answers to these speculative questions, let alone to claim that I know some set of answers to be true. It seems more truthful to say that I do not know what moral obligation is, at least in the restricted sense that would involve confident and justified assent to a perfectly complete essentialist definition. I am prepared to say that some version of the social theory of obligation is correct. Obligations are requirements that arise in social relationships. Whether they are morally binding has something to do with whether the underlying social relationship is good or excellent in certain respects that would be hard to specify in full. We know what some paradigmatic examples of defective relationships look like—the relationship between master and slave being one and that between a conqueror and a conquered people being another. Generalizing, I can say that a social relationship becomes defective in the relevant sense in so far as it involves domination analogous to relationships of these types. This modest bit of theorizing already gives me a reasonably firm grip on the notion of a morally binding obligation, and it orients my future inquiry into obligation in a helpful way. But it does not provide the sort of essentialist definition that Socrates was looking for, because that sort of definition would require us, in the present case, to specify *all* sorts of possible defect in a social relationship that might render a social requirement less than fully binding on a dominated party and to specify precisely *how* binding, if at all, a requirement would be if it arises in a relationship defective in this or that respect or to this or that degree. It is wise not to claim too much knowledge about such matters—more knowledge than a finite human being can muster in the present stage of ethical inquiry.

For Adams it is the task of 'the metaphysical part of ethical theory' to supply something much closer to a full-fledged essentialist definition. Metaphysical inquiry into the nature of moral obligation or of the good, he says, is analogous to scientific inquiry into the nature of water. 'But the

nature of water is to be discovered in the water and not in our concepts'.[23] So we should not be looking in our concepts to discover the nature of moral obligation. Where, then, should we be looking and by what means? What is supposed to be analogous here to looking *in the water*? For that matter, is it clear that human beings did not know the *nature* of water before the physical sciences informed them of what *constitutes* water?

Mark Johnston argues that what chemists discovered in the water was what water *is made of*—its constitution—not what it *is*. Water, according to Johnston, is a manifest kind, not a natural kind. As a manifest kind, what it *is* must be identified in terms of its manifest properties (the properties that manifest themselves to human beings under ordinary circumstances). It is not to be identified with H_2O. Thus, chemical inquiry has not revealed the nature of the manifest kind, but rather the constituent parts that go together to make up instances of the manifest kind. Chemistry 'does this by showing how the chemical kinds in question account for the causal profile of (instances of) the manifest kinds in question'. But doing this, Johnston insists, does not involve reducing manifest kinds to natural kinds. Failure to understand what is wrong with such reductions is a major source of the upsurge of scientism in contemporary philosophy.[24]

Now, if we accept Johnston's understanding of the water example, what becomes of the metaphysics of morals? If the morally obligatory is analogous to a manifest kind, its nature would be something implicitly grasped by all human beings who (rightly) take themselves and others to be morally obligated in various ways. It would not be identifiable in terms distinct from the vocabulary of ordinary ethical deliberation and accountability. We would look for the nature of moral obligation, then, *in these practices and in the lives and relationships of the persons engaging in them, but without being tempted to reduce obligation to a matter of biological, psychological, or social fact*.[25]

Johnston rejects the reduction of manifest to natural kinds that is characteristic of scientism. The parallel in ethics would be to reject the

[23] *Finite and Infinite Goods*, 15.

[24] Mark Johnston, 'Manifest Kinds', *Journal of Philosophy,* 94 (1997), 564–83. See also id., 'Constitution is Not Identity', *Mind*, 91 (1992), 83–102. How much would Soames's history of analytic philosophy have to be rewritten, I wonder, if the distinction between constitution and identity came to be viewed as equal in importance to some of the distinctions drawn in Kripke's *Naming and Necessity*?

[25] In saying this, I am neither endorsing the theory of value Johnston has developed in other articles, nor claiming that Johnston would endorse the use I am making here of his treatment of manifest kinds and the distinction between identity and constitution.

reduction of ethical properties to biological, psychological, or sociological facts. Biology, psychology, and sociology cannot tell us, in a non-normative vocabulary, the nature of obligation. They can tell us how a species of norm users evolved, how a member of this species typically advances from one to another stage of moral development, and how a community's practices of deliberation and accountability function. In telling us such things, these disciplines shed light on ethics in a way analogous to the way in which chemistry sheds light on water. The causal stories they tell can be highly informative, but they do not have a direct bearing on what it is to be morally obligated, nor do they tell us much about what our actual obligations are. To discover such things, we have to reflect, from a normatively committed point of view, on our experiences in different sorts of relationships.

We all stand in social relationships that give rise to requirements of various kinds. When the underlying relationships are sound, we ordinarily take the requirements arising within them to be binding. Other things being equal, we feel bound by those requirements. The experience of being mutually obliged is intertwined with the practice of holding one another responsible for our actions within the context of our relationships. When we breach those relationships, we typically experience guilt. When another person breaches one of those relationships, we often feel anger. We are also able, however, to recognize defects in our relationships. Some of these defects weaken the normative grip of requirements that have been imposed on us. When we are in a severely defective relationship, and we are the disadvantaged party, we *should not* feel guilty if we violate such requirements. And if our thinking has not been unduly distorted by the defective relationship, we *do not* feel guilty. By reflecting on such experiences, we learn something about the nature of obligation and the difference between genuinely binding obligations and social requirements that are merely imposed by the dominant on the dominated.

Like Socrates, I know enough about moral obligation to initiate my inquiry into it, and to earn entitlement to some judgments about what it involves, but not enough to claim fully explicit discursive mastery of its essence or nature. By finding myself in relationships that impose requirements and by participating in the practices of deliberation and mutual accountability, I have learned a good deal about moral obligation. Making that knowledge explicit and rectifying my use of the concept are legitimate philosophical tasks, but they do not lift me out of the relationships

and practices I am reflecting on. The point of view of the inquiry remains a normatively committed one. And the inquiry is reflexive; the object of inquiry is the life being led by the inquirer.

What happens when the philosopher who models moral metaphysics on science turns to theology for explanations of what moral obligation and excellence are? It is clear why Adams finds theological explanations attractive. He is 'inclined to think that the representation of anything rich enough to be a perfect standard of excellence must depend on its actual exemplification'. Anything less, he says, would leave us without an 'objective standard of excellence'.[26] Construing excellence and moral obligation as objective in the strongest possible sense is the primary motive for Adams's introduction of theology into the metaphysics of morals.

It is noteworthy that the unmodified version of divine command theory, which appeals simply to what God commands, without using a normative term to characterize God, is structurally similar to scientistic reductions of moral obligation to a matter of objective fact. By explicitly characterizing God as loving, Adams's account of obligation remains within the normative sphere, and in this respect is unlike a scientistic reduction. It seems wise, all things considered, to accept that the nature of moral obligation is going to remain bound up with the social practices and relationships in which we become intimately familiar with it. Of course, if God exists, enters into covenants with human beings, and issues commands to them, then God is one of the persons involved in the relevant social practices, and God's contribution to those practices will have to be considered in any complete account of them. If God does not exist, however, then the only persons to be taken into account will be our fellow human beings. Either way, the practices and the social relationships they involve would be the place to look when reflecting on the nature of moral obligation.

Adams does not make clear why securing the conclusion that excellence and moral obligation are objective *in the strongest sense* should be treated as a desideratum for theory-construction in moral philosophy, instead of as a standing temptation to endorse some form of reductionism. It seems to me that the prior questions ought to be: first, given an honest appraisal of evaluative discourse and deliberation, what sort of objectivity, if any, appears to be achievable in this part of our lives? and, second, why should

[26] *Finite and Infinite Goods*, 44–5.

philosophy be expected to provide a justification or explanation of our evaluative practices that appeals to something more reassuring and grand than their evident value to those participating in them?

If Adams is simply setting out from antecedently held theistic premises, and trying to account for excellence and moral obligation in terms of those premises, he is engaging in a perfectly respectable exercise in expressive rationality. It is quite possible that he is entitled to his theological commitments. He is certainly entitled to make those commitments explicit and explore their implications for an understanding of excellence and obligation. Philosophy makes progress, it seems to me, when we all make our commitments explicit, explore their implications, and expose them to critical questioning. This is what the Socratic model of philosophical inquiry requires of us, and Adams is a master practitioner of the art. But, in so far as Adams argues *to* theistic conclusions by appealing to their value in securing maximal objectivity for evaluative claims, or models moral philosophy on scientific explanation, I find the enterprise dubious.[27]

[27] I wish to thank Mark Johnston, Nicholas Wolterstorff, and above all Bob Adams, among other friends and colleagues, for helpful comments on the much longer version of this paper that was discussed in the Yale symposium in Bob Adams's honor in April 2005. Thanks also go to Sam Newlands for insightful comments on the penultimate draft of this shorter version.

12

The Grasshopper, Aristotle, Bob Adams, and Me

SHELLY KAGAN

The Grasshopper of my title is, of course, the insect made famous by Aesop in his fable concerning the Grasshopper and the Ant—the grasshopper who praises the virtues of idleness and play, and who perishes with the coming of winter because of his prior refusal to engage in the instrumentally necessary but decidedly unenjoyable task of gathering food. I focus, however, not upon Aesop's report concerning the Grasshopper's views, but rather on the less well-known, but philosophically richer, report provided by Bernard Suits, in his amazing book, *The Grasshopper: Games, Life, and Utopia*.[1]

In that book the Grasshopper puts forward and defends a definition of games, which he helpfully summarizes with the slogan that playing a game is 'the voluntary attempt to overcome unnecessary obstacles'.[2] The basic idea is this: when I play a game, I am trying to accomplish some goal; but I am not trying to accomplish that goal by the most efficient means available. Indeed, I deliberately accept restrictions on how the goal is to be achieved. For example, in golf, I am trying to get balls in holes, but I do not allow myself to simply pick the ball up and place it in the hole—rather I restrict myself to getting the ball in the hole by means of hitting it with the right kind of stick. And in mountain climbing, I am trying to get to the top of the mountain, but I do not allow myself to take the helicopter to the summit; rather, I restrict myself to getting there by climbing up along the side, using only the relevantly permissible equipment. Why do I accept these unnecessary obstacles? Because doing this allows me to engage in an

[1] *The Grasshopper: Games, Life, and Utopia* (Toronto: University of Toronto Press, 1978); citations to this edition are followed [in square brackets] by corresponding citations to the 2nd edition (Peterborough, Ont.: Broadview, 2005).

[2] *Grasshopper*, 41 [55].

activity that cannot otherwise be performed: playing the game. I accept the restrictions so as to be able to play the game. If that's my reason (or at least one of my reasons), then I have what the Grasshopper calls the 'lusory attitude'. Without it, I may look just like someone playing the game, but I am in fact simply going through the motions.

I offer this brief rendering of the Grasshopper's account of games, because it plays a central role in what I really want to discuss—some utopian speculations that the Grasshopper puts forward as well. What I really want to ask is this: just what would we do in Utopia?

In thinking about Utopia, I want to let our imaginations run wild and assume—again, following the Grasshopper's lead—that *all* technological limitations have been overcome. Computers or friendly spirits or magic dust can instantly provide whatever it is we might want. Thus, there is no *need* to do anything so as to achieve something else: no need to exercise to preserve health, no need to eat to acquire nutrition, no need to work to attain clothing (or books, or housing), no need to study or investigate to attain knowledge. There is no scarcity of any sort, so no need for accomplishments of any kind. Problems of interpersonal conflict (most of which turn on problems of scarcity, in any event) have been solved, and so there is no need for government;[3] problems of science and of philosophy have all been answered, so there is no need for scientists or even (gasp!) philosophers.

So what I want to know is this: what would we do in Utopia?

It would be natural to think that this question comes to the same thing as asking: what activities are intrinsically valuable? After all, what the concept of Utopia (carried to its logical limit) invites us to do is to imagine a world where we have eliminated the necessity of performing an act simply because of its instrumental value. If we do something in Utopia it isn't because we need to do it so as to get something else. Instead, if we do something, it must be because we take that activity to be intrinsically valuable—valuable for its own sake, not merely as a means to something else. That is to say, in Utopia there is no need to engage in instrumentally valuable activity (since anything we want to produce thereby can be achieved effortlessly instead).

[3] Arguably, many of the moral virtues—and, more generally, much of the need for morality—will have been eliminated in Utopia as well. If there is no interpersonal conflict, for example, there is no need for justice; if there is no disease, hunger, or poverty, there is little or no need for compassion, beneficence, or self-sacrifice. But the issue is complex and I won't try to pursue it here. (For example, have we also eliminated *mortality* in Utopia? If not, there may still be a place for compassion, among other things.)

But even in Utopia, presumably, we will want to engage in intrinsically valuable activities. So it is natural to think that when we ask what we would do in Utopia we are asking for an account of which sorts of activities possess intrinsic value. The many things that we do in the actual world but which we would not perform in Utopia must be merely of instrumental value, rather than intrinsic value.

As it happens, I believe that this natural inference is mistaken. It assumes that the intrinsic value of intrinsically valuable goods cannot be grounded in part on their usefulness. In particular, it assumes that if some activity is instrumentally valuable, then that very fact cannot contribute to the activity's also having intrinsic value. And while this assumption is very widely held, I believe it is mistaken. I think we should recognize the possibility of intrinsically valuable instrumental value.[4]

This means that there could well be activities in which we currently engage, activities in which we engage because we must—given current technological limitations—activities that are instrumentally valuable, and yet in part precisely because of that fact are also intrinsically valuable. If there were activities of this sort, however, then although in our actual world they would have intrinsic value, in Utopia they would *lack* intrinsic value. For in Utopia these activities would no longer have significant instrumental value, and so would lack an essential part of what currently grounds their intrinsic value. If I am right about this, then it turns out that the list of intrinsically valuable activities is broader than the list of activities we would engage in within Utopia. Oddly enough, Utopia will make certain sorts of intrinsically valuable activities impossible (or, more accurately, will strip them of their intrinsic value).

So when we ask what we will do in Utopia, this isn't quite the same thing as asking for a complete list of intrinsically valuable activities. We will have put aside those activities whose intrinsic value is based on necessity.

Does this mean that we should mourn the passing from Utopia of the no longer intrinsically valuable activities? This threatens to be a paradoxical result, suggesting that the 'ideal' of human existence (as the Grasshopper calls it) would itself be impoverished and lacking. But I am not sure how best to avoid this unhappy conclusion.

[4] I have argued for this conclusion in 'Rethinking Intrinsic Value', *Journal of Ethics*, 2 (1998), 277–97.

One possibility, of course, would be to simply reject my claim that intrinsic value can be grounded, in part, in necessity. Then we can insist that all genuinely intrinsically valuable activities remain possible in Utopia. This may seem the obviously sane solution to most of you, but unsurprisingly, it does not seem right to me.

An alternative proposal is suggested by some remarks made by Bob Adams in his wonderful *Finite and Infinite Goods*.[5] This is perhaps a good place for me to admit that I have a rather difficult time trying to read this book. Whenever I open it, I find my mind racing, overflowing, jumping off into new lines of thought. I am, rather literally, inspired. Questions occur to me that I haven't previously entertained, and I excitedly go off exploring philosophical possibilities that have been suggested to me by this or that remark in the text. When I 'come to', several minutes later, I see that I am several pages further along in the book, but I am not at all confident I have actually been reading. This is by way of confessing that although I will occasionally refer to a few of Bob's views, I won't be doing this in an appropriately scholarly way; I haven't yet been able to read the book properly enough to do that.

Anyway, the particular remark I have in mind here is the simple observation that there are many goods, and no single human life may be able to incorporate all of them.[6] Perhaps we could take this idea and accept it writ large (as Bob does in his discussion of liberalism[7]), holding that no single society—not even utopian ones—could contain all intrinsically valuable activities. If some intrinsically valuable activities have become impossible under Utopia, so be it. So long as new and better goods are available in Utopia, it may be worth the cost.

But this leads us to ask anew: just what will we do in Utopia?

The Grasshopper's own answer to this question is straightforward: we will play games.[8] After all, there is no need for me to do anything at all in Utopia. Whatever it is that I am trying to bring about, it could have been attained instantly and effortlessly (say, through the magic dust). If I nonetheless persist in trying to bring about some goal, I am deliberately doing this by less efficient rather than more efficient means. I am trying to see if I can

[5] Robert M. Adams, *Finite and Infinite Goods* (Oxford: Oxford University Press, 1999).

[6] Where did Bob say this? I'm not sure! (Remember, I've barely read the book.) But the idea is suggested, at least, by comments on pages 57 and 292 of *Finite and Infinite Goods*.

[7] *Finite and Infinite Goods*, 334.

[8] See, especially, *Grasshopper*, chapters 1 and 15.

'do it myself'—using my bare hands, or these primitive tools, or these artificially limited means. In short, whatever it is I am doing, I am setting myself some unnecessary obstacles, and voluntarily trying to overcome them. Whatever it is I am doing, I am playing a game.

These won't necessarily be the sports, board games, and pastimes we are familiar with from ordinary life, but they will be games nonetheless. I might, for example, try to build a house using carpentry tools and lumber; but since I am voluntarily making it harder on myself (why not just order a house from the computer?) this is just playing the 'house-building game'.[9] Similarly, I might decide to try to solve some math problem myself—or to think through some philosophical conundrum on my own—but since I am voluntarily making it harder on myself (why not just look at the answer page, with its lucid explanation?) this is just playing the 'math game' or the 'philosophy game'.

It might be objected that I cannot possibly come to understand the math, say, without first having struggled to master it. But this reply, I think, fails to take seriously the assumption that in Utopia *all* technical problems have been solved. I can—we are to assume—not just learn the answer, but also come to have a complete and deep grasp of the underlying mathematics simply by taking a pill (or indeed, simply wishing it to be so). If, then, I nonetheless insist on studying mathematics, I am playing a game: trying to see how much I can learn 'the old fashioned way'.

The Grasshopper's claim then is that in Utopia what we will do, and all we will do, is play games.[10]

One might reasonably worry whether this is enough to sustain us. Can we view our lives as having sufficient value and significance, if all we are doing is playing games? The Grasshopper himself has a nightmarish vision in which the Utopians disappear, one by one, as he reveals to them that all they are doing is playing games. A life devoted to game playing hardly seems worth living.

Accepting this unpleasant conclusion does not require us to hold that game playing is not, in fact, an intrinsically valuable activity. We might

[9] Cf. *Grasshopper* 174 [156–7].

[10] Actually, the Grasshopper claims that with games as the centerpiece of utopian life, other activities—activities focused on games, as it were—become intelligible as well (see 176 [158]). It seems to me, however, that the logic of the Grasshopper's argument should nonetheless support the claim that even these further activities are themselves games (e.g., if one tries to produce by oneself artistic renderings of game playing, rather than simply 'ordering some up', one is still playing a game).

well agree that there is indeed intrinsic value in playing (well-crafted) games,[11] and yet still worry that in and of themselves games are not a source of sufficient value to form the essence of a good life.

Suppose that you are with me—and with the Grasshopper, at least in his darker moments—in suspecting that game playing is not a rich enough diet to make life in Utopia worth living. That's a problem, because what we set out to do, recall, was to investigate the ideal of human existence—where we have solved all the technological problems, and so can have whatever intrinsic goods there are to be had (subject only to this very technological assumption, that we can have whatever we want effortlessly). And now we find ourselves worrying that the ideal of human existence may not be all that ideal after all.

Clearly, this takes us beyond the Adamsesque thought that you can't have it all—that is, everything worth having—in a single life, or a single society. Here we are saying you can't have a sufficiently good life in Utopia to be worth living. That seems decidedly odd.

Of course, it could be maintained that we erred in thinking that it was indeed a move toward *Utopia* to imagine unlimited technological solutions (so that there was nothing we needed to do). It would hardly be surprising if the imposition of a sufficiently unattractive feature upon society as a whole leaves us with only the possibility of impoverished lives.

But although this reply is coherent, I find it hard to take seriously. I find it hard to affirm the claim—and to stick with the claim—that the technological assumption is a mistaken move of this sort.

First off, in pursuing the idea of an *ideal* life it does seem reasonable to explore it under ideal conditions. For even if it is also of interest to think about the best kind of life available given certain features that we take to be undesirable, it seems plausible to think that the very *best* kind of life would be one in which those undesirable features are eliminated. This would, indeed, represent a conception of the *ideal* of human life, however unrealistic it might be. But then, second, it does seem as though our various technological limitations are undesirable features of our current situation. Admittedly, we may be driven to deny this, by virtue of the very utopian reflections in which we are engaged, but when considered directly, at any

[11] As Gwendolyn Bradford argues in 'Kudos for Ludus: Game Playing and Value Theory', *Noesis*, VI (available online at <www.chass.utoronto.ca/pcu/noesis/issue_vi/noesis_vi_3.html>).

rate, it seems virtually self-evident that our technological limitations are undesirable features. One need only think of the constant human striving to overcome them, a feature that would be well-nigh inexplicable if they were in fact overall *desirable* features of the human condition.[12] Yet if technological limitations are in point of fact (overall) undesirable, then they are features which are appropriately imagined eliminated when thinking about the ideal form of human existence.

And that brings us back to the thought that in Utopia there is nothing to do except play games, and that this may not be nearly enough.

So what else, if anything, is there to do? I throw this question out in a genuine spirit of dialogue. I am not confident of the answer, nor in fact am I confident about much of anything I am working through here today. I have a few ideas that I am intrigued by, which I will share, but I don't yet have anything close to considered views.

One move that may be helpful here would be to distinguish in a very broad way between production and consumption. In effect, what the technological assumption does is rob our productive behavior of much of its point, since there is no longer any need for us to produce anything. That appears to leave us only with the option of productive behavior that is unnecessary, and indeed harder than it needs to be—behavior that produces a goal through deliberately inefficient means: game playing. But even if that is right with regard to productive behavior, it may still leave us with various types of consumptive behavior in place. In particular, then, there may still be an array of intrinsic goods waiting for our consumption, even in Utopia. Indeed, there may be finer, and greater, intrinsic goods available for our consumption in Utopia than are available now.

Thus, there may be very little for us to 'do' in Utopia—other than playing games—if by talk of 'doing' we mean to focus exclusively on

[12] Bill FitzPatrick has suggested to me that since so many of the activities that currently give our lives meaning involve the overcoming of one or another kind of involuntary obstacle, the ideal of human life will indeed involve such obstacles as well—with an 'optimal' level of obstacles. But I think there is something troubling and unstable about the suggestion that even in the ideal life we would still strive to overcome involuntary obstacles: since success in overcoming these obstacles would leave us with 'too little' to overcome (less than the optimal level), we would also have to hope we don't succeed in overcoming them! (It may also be worth pointing out that FitzPatrick's suggestion implicitly assumes the existence of intrinsically valuable instrumental value, since the intrinsic value of the relevant meaningful activities depends, in part, on their role in overcoming obstacles. Of course, I don't myself take that fact to constitute any kind of objection to the proposal; but others may.)

productive activities. But we should not neglect to consider the various consumptive activities.

This talk of consumption may be somewhat ill-advised, suggesting as it does mere material consumption. It tempts us to imagine an array of finer and finer plasma TVs, or elegant houses, or exquisite foods (or who knows what technological marvel). To be sure, even this may be the source of some intrinsic value. The consumption of these various material goods would not be necessary, but would presumably still be pleasurable—perhaps even extraordinarily pleasurable. But I am surely not alone in worrying that this too may not be a rich enough diet to sustain us. A life of food, drugs, games, and TV sets need not be dismissed as altogether empty; but it still seems altogether too shallow to qualify as the ideal of human existence.

Another idea of Bob's seems helpful at this point. Bob suggests that well-being consists in the enjoyment of the excellent.[13] I have a less elegant mind than Bob's, so I hope I am not too far amiss in glossing this as the claim that well-being consists in taking pleasure in the possession and consumption of significant intrinsic goods. This raises two points of interest. First, and most obviously, if one takes pleasure in something that isn't genuinely good, this may make little or no contribution to one's level of well-being. But second, and less obviously, if the object of one's enjoyment is good but not significantly good, it will fall short of excellence and so, again, make little or no contribution to well-being. Both of these worries seem germane in thinking about the utopian life we have described so far. Game playing and material consumption may not be without value, but they seem to fall sufficiently short of enjoyment of the excellent to justify our concern that such a life cannot truly constitute human well-being, let alone the *ideal* of human existence.

But as I noted, I intend the category of consumption to be construed quite broadly, and we might hope to find greater goods to consume in Utopia. Here is an example. Arguably, knowledge is an intrinsic good. Given the technological assumption, of course, there is no need in Utopia for study, inquiry, or investigation to gain knowledge: everything there is to know (or, at least, everything that can be known by humans) can be known instantly, effortlessly. So the pursuit of knowledge has no place, unless, it seems, as part of an 'inquiry game'. But for all that, the 'consumption' of

[13] *Finite and Infinite Goods*, chapter 3, section 2.

knowledge may still be good. That is, there may be no great value in our *striving* to know, but there may for all that be value in our *knowing*.

Of course, it would not be plausible to suggest that all instances of knowledge are equally valuable. If I happen to know the average daily rainfall in Bangkok in February 1993, there isn't likely to be much value in that. But it seems plausible to suggest that knowledge is more valuable when the truths known are themselves more *important*—when they are more significant, deep, and profound.

Suppose, then, that in Utopia I am able to know the fundamental laws of nature—not just recite them, in the way that I may be able to rattle off the laws of Newtonian Mechanics, or Heisenberg's Uncertainty Principle, but to fully grasp them: I understand what the laws mean, and I can see just how they suffice to generate and explain the astonishing array of empirical phenomena. There may well be significant value in simply understanding all of this. If Utopia could provide us with that, then that might well constitute something significant to do with our lives: we could contemplate the fundamental truths of the universe.

Or we might go further. Suppose that there is a creator. Indeed, suppose that we could grasp enough about the divine nature to comprehend something significant about God's power and infinite goodness, grasp enough to see not only *that* God created the universe but also something about God's purposes in doing so, enough to grasp our place in a universe brought about and sustained by God's love. That would be significant, deep, and profound knowledge indeed. Contemplating God and God's plan for the universe might well be something that would be sufficiently worthy of our time.

No doubt in thinking about this last possibility we have moved beyond any mere Utopia, at least insofar as it is unlikely that any merely technological fix could bring us to this point. But insofar as what actually drives our discussion is a quest for a satisfying and perhaps inspiring account of the ideal of human existence, this theological possibility does not seem out of place. What would we do in Utopia? Perhaps we would contemplate the divine. Indeed, if we could, we would partake of the beatific vision (though no doubt on certain views it would take divine grace to find ourselves in a position to do so).

I find this theological possibility attractive insofar as it seems to have sufficient weight to provide part of a satisfying account of the ideal of human

existence. I imagine that Bob may find it attractive too, though he mentions the beatific vision only once, and then only in passing.[14] But despite this point of agreement, there is probably an important difference between us. I suspect—though to be sure, I do not know—that Bob actually believes in the possibility of this kind of mystical vision ('in this life or the next', as he puts it); I do not. (The disagreement isn't so much epistemological as metaphysical; I just don't have the requisite religious beliefs.)

Be that as it may, this discussion puts one in mind of Aristotle's claim that the best kind of life for humans would consist solely in philosophical contemplation—in particular, I take it, contemplation of the divine.[15] Previously I have always found this the sort of absurd claim that only a philosopher could make, a claim hopelessly and implausibly glorifying philosophy to the exclusion of everything else. But now it seems to me possible to understand the Aristotelian claim as an instance of this same general suggestion concerning the ideal of human existence. And now, I must confess, it doesn't strike me as quite so absurd. (Just like a philosopher to find such an absurd claim worth taking seriously!)

Indeed, the crucial point is to emphasize that the claim might make sense if taken as a thesis about the *ideal* of human existence. The Aristotelian claim seems most absurd insofar as it seems to overlook the variety of intrinsically good activities that we appropriately engage in as part of our ordinary, everyday lives. But our present condition is far from ideal, and as such it is open to us to insist that various activities are intrinsically valuable today—given the technological limitations of our current condition—and yet for all that would form no part of the ideal. I honestly don't know whether this is, in fact, at all faithful to the sorts of considerations actually moving Aristotle (believe me, I am no serious student of Aristotle); but it seems to me at least one way of making some sense of these claims. (Note, in particular, how it allows Aristotle to maintain that the practical or political life is a kind of second best—intrinsically good, but no part of the ideal of human existence.)

There is a somewhat different direction in which we might try to develop the proposal that at least one of the important things we will do in Utopia is to know things. Suppose we back away from the theological speculations of the last few paragraphs, and return to the idea that what we will contemplate

[14] *Finite and Infinite Goods*, 23.

[15] See, especially, *Nicomachean Ethics*, Book X, chapters 7–8.

in Utopia are the fundamental truths of the universe. Among these truths, perhaps, are the following: the world is itself intrinsically valuable, and it has produced creatures like us (perhaps via various naturalistic processes), creatures who themselves possess intrinsic value. Indeed, the world may itself have intrinsic value precisely by virtue of the fact that it produces these intrinsically valuable creatures (this would be a nice illustration of an intrinsically valuable instrumental value), and the intrinsically valuable creatures produced may be intrinsically valuable in part precisely because of their ability to come to recognize and know all of this. Suppose all of this were true, then perhaps one of the most important things we would contemplate in Utopia would be the very fact that the universe is valuable precisely because of having created creatures like us, capable of contemplating this very fact.

This strikes me as being a rather Hegelian suggestion. At least I take it to be broadly Hegelian in spirit, though I can't really say (I am even less of a student of Hegel than I am of Aristotle). Perhaps, then, this talk should actually be entitled 'The Grasshopper, Aristotle, Bob Adams, *Hegel* and Me'! At any rate, I mention it because although I am far from confident that it is true, it seems to me that something like it might well be true.

But even if one is unprepared to travel these Hegelian circles, or for that matter to put faith in the earlier theological speculations, there are presumably some fundamental truths concerning the universe which we are capable of coming to know. And so it remains possible to hold that whatever these truths might be, in Utopia one of the things we will do is to contemplate them.

Nor is this the only significant intrinsic good which we are capable of consuming in Utopia. Up to this point I have said almost nothing about art and beauty. Now the Grasshopper worried that there might be no place for art in Utopia. His thought was not only the by now familiar one that there was no point in producing art (except as a game) but rather the deeper worry that given the elimination of suffering, frustration, and conflict in Utopia most of the subject matter of great art will have been eliminated as well.[16] (No wars, for example, means no masterpieces depicting the horrors of war.) If true, this is yet another odd implication of the concept of Utopia.

[16] This is surely too quick, since it is, for example, unclear whether *death* has been eliminated from Utopia (cf. n. 3, above). But let it pass.

Of course, this might just be another occasion for reminding ourselves of the Adamsesque thought that even in Utopia certain goods must be done without. Indeed, the very production of Utopia may destroy the possibility of some undeniably significant goods.[17]

But instead of pursuing this question here, I will simply note that even if certain forms of art require suffering and strife for their subject matter, it would not be plausible to claim that all art is like this. If, for example, as seems plausible to me, moments of great joy need not involve the overcoming of obstacles, then there could presumably be great art depicting such moments, even in Utopia. And in any event the possibility of abstract, nonrepresentational art remains as well. What's more, there also remains *natural* beauty, which might be the subject of still more art—as well, of course, as being the direct object of aesthetic appreciation. And, of course, there is also the more abstract beauty to be found in mathematics as well, or in certain fundamental laws.

Still other forms of art—such as music—seem like they should be available in Utopia as well. Without trying to take on the difficult and troubling issue of the connection between emotion and music, let me just suggest that even if some music will not be available (or will not be intelligible) in the absence of struggle and failure and loss, other forms of music—including much great music—should still be possible.

All of this provides still more goods for consumption. In Utopia we will enjoy the beauty of the natural world—the Grand Canyon, the Redwoods, and the Himalayas, as well as the flash of a cardinal or a hummingbird—or the beauty of fractals, or quantum mechanics, or Monet's water lilies, or Mondrian's squares, and we will listen to music (if not much Beethoven, perhaps more Mozart).

We could, presumably, combine the consumption of knowledge and beauty, in ways that I have already hinted at. The divine may not only be profound, but beautiful; and the same could be true—less theologically—for the laws of nature. But I see no reason to think that consumption of the beautiful should be limited to the most grand. There should still be

[17] On the other hand, given the assumption of unlimited technology, can't we simply 'order up' some artistic depictions of suffering and conflict (even if these won't themselves exist in Utopia), and then use these works (or other technology) to learn about the nature of suffering? If so, we might both possess and appreciate such works of art after all, even in Utopia.

a place in the ideal of human existence for listening to an occasional Sousa march.

So among the things that we will do in Utopia, besides playing games, are knowing and admiring. These may not be activities in the normal sense of the word, but they may make up worthy ways to live even in Utopia.

I have been suggesting that we may find something to do in Utopia after all, once we recall that we can consume as well as produce. Even if the only productive activity worth engaging in within Utopia is to play games, there may still be other things to *do*. But I would not want to leave it here, implying that the Grasshopper is at least correct with regard to the productive side of the divide. On the contrary, even with regard to productive activity, I believe, there is more to do in Utopia than to play games.

For concreteness, let's consider the case of producing a work of art. Suppose I paint a picture of a rose. By hypothesis, of course, there are easier ways to get such a painting in Utopia; I could just order one up. So if I choose, nonetheless, to produce a painting in the traditional fashion—working with a canvas, paints, and a brush—there is certainly a sense in which I am choosing less efficient means to achieving my goal, rather than more efficient means. And so—insists the Grasshopper—I am simply playing a game: the 'art-making' game.

But at this point we should remember the Grasshopper's earlier claim that an essential part of playing a game is having the lusory attitude. You must be voluntarily trying to overcome unnecessary obstacles so as to be able to participate in the very activity that thereby becomes possible (that is, becomes possible through the undertaking of unnecessary obstacles). In effect, you must be playing by the rules so as to be able to play a game. If you act for different reasons (that is, reasons not including this one) then you are not actually playing a game at all, even though it may well look like you are to the outside world.

Well, what other reasons might one have for trying to make art the old fashioned way, if not the desire to play a game?

The natural suggestion to make, I think, is this: in painting a rose I am being creative; in ordering up a painting, I am not. (This is not to deny that one could be creative through sufficiently creative specification of the product desired—think here of a director of a movie—it is only to insist that if all you are doing is ordering up a painting, then you are not being

creative.) So perhaps my reason for doing the painting myself is that I believe, plausibly, that creative activity is itself intrinsically valuable, indeed it can be excellent.

Here another suggestion of Bob's comes to mind. Bob suggests that something is excellent if it resembles God.[18] Presumably then Bob would agree that creative acts can resemble the creativity of God, and thus be excellent. And there is no obvious reason to believe that creativity ceases to be excellent simply because it isn't *necessary*.

Now here too, I am not quite inclined to agree with Bob completely. I would rather turn the idea around. Bob thinks that being creative is excellent because it resembles God. I believe, rather, that it might be closer to the truth to claim that the reason we ascribe creativity to God is because we recognize its excellence. But this is a debate for another occasion. Here the point is just this: if creative activity is excellent, then I can choose to engage in the production of art, not because this is the only way to play the art game, but because this is one good way to be creative. In any event, to be creative I must do something: I must *produce* the work of art.

Thus there are reasons to produce a work of art that, so far as I can see, have nothing to do with game playing. The efficiency or the inefficiency of the means are, so far as I can see, beside the point. At any rate, I am not, in the relevant sense, voluntarily trying to overcome unnecessary obstacles. Were there better ways to be creative, I might well choose them. If the means are inefficient, I am willing to put up with this, but the inefficiency is not, in and of itself, part of the attraction. In short, I lack the lusory attitude. I am not playing a game.

Once we see this point, it seems likely that we will find it replicated in other productive activities as well. Suppose I tell my wife that I love her. Why am I doing this? The Grasshopper insists that, given the assumption of unlimited utopian technology, there must be a way for my wife to know about my love for her without any effort on my part. So if I insist on letting her know in *this* way—especially given the attendant dangers of miscommunication—I must be playing another game, perhaps the 'communication' game.

But that seems wrong. I am not merely trying to make it be the case that my wife knows I love her. I am trying to *express* my love. That is not

[18] *Finite and Infinite Goods*, chapter 1, section 3.

something I do if the communication of the message does not go 'through' me. The message can be received, presumably, even if it is said by another on my behalf; but it is not then said by me. And there is, arguably, value in the deliverance of the message by me. There is value in my *revealing* my love, and not only value in her knowing about it. Here too, the efficiency or inefficiency of my means is beside the point. I am willing to put up with the inefficiency, but it is not, in and of itself, part of my reason for acting. I lack the lusory attitude. I am not playing a game.

More strictly, I *need* not be playing a game. The Grasshopper, recall, does not insist that the lusory attitude be my only reason for taking on unnecessary obstacles. So long as it is *part* of my reason for doing this, I am playing a game. So we can allow that I might in fact be playing a game after all, if the desire to do this is indeed part of my reason for acting as I do.

But it would not then be true to say that *all* I am doing here is playing a game. It need not be my only reason for engaging in my various productive activities. It need not even be my main reason.

It is less clear to me whether we can extend this same basic line of thought to cover *inquiry* as well. Suppose that in Utopia I try to figure out something for myself. The Grasshopper insists that, given the possibility of attaining the relevant knowledge instantly and effortlessly, I must be playing a game. What other reason could I have for doing it myself in this way?

This reminds me of the famous passage from Gotthold Lessing:

> If God held all truth enclosed in his right hand, and in his left hand the one and only ever-striving drive for truth, even with the corollary of erring forever and ever, and if he were to say to me: Choose!—I would humbly fall down to him at his left hand and say: Father, give! Pure truth is indeed only for you alone![19]

The Grasshopper agrees, of course, that there is value in Lessing's choice. He only insists that, whether or not Lessing realizes it, in making this choice he must be choosing to play the inquiry game. This is, no doubt, a fine game to play, but a game it must be.

Is that right? Or might there be other reasons for engaging in inquiry in Utopia? Is there something intrinsically valuable about the search for knowledge itself, as opposed to the possession of knowledge—or more precisely, is there something intrinsically valuable about the search for

[19] As quoted by Kierkegaard in *Concluding Unscientific Postscript*.

knowledge in a world in which such a search is no longer a necessary means to acquiring knowledge? Is there something excellent about inquiry per se? If there is, I have to confess, I cannot yet see what it is. (Essay question: Does inquiry resemble God? If so, how? Comment: It may be helpful to consider the rabbinic tradition that portrays God as himself engaging in the study of Torah.)

But even if it should turn out that inquiry is not one of the productive activities in which we would engage in Utopia, except as part of a game, it still seems correct to suggest that the same is not true for other productive activities. There are other things to do—in the productive sense—besides playing games. I would particularly want to emphasize the thought that even in Utopia there will still be intrinsic value in maintaining a wide range of meaningful personal relations.

The basic idea here is one that has already been suggested by my discussion of the value of telling my wife that I love her. What is of value in such a case isn't merely the state produced (that is, my wife's knowing of my love) but the *production* of such a state by me (my *telling* her). Something similar presumably holds true for a variety of personal relations.

We might put the point a slightly different way: there is intrinsic value in relationships of the right sort, but the value to be had here is not nearly exhausted by the simple fact of *standing* in those relationships; there is also value in *relating*. Thus, even if we can make sense of the idea of there being more efficient ways of coming to stand in the relevant relationship, or of preserving the fact that one stands in that relationship, that would not yet give us reason to forgo the *acts* of relating to others in the relevant ways (even if these were less efficient ways of producing the relevant state).

But there is still more to say: in many cases it doesn't seem implausible to suggest that a good deal of what it is to *stand* in the relevant relationship just is to relate to one another in the relevant ways (in a sufficiently regular and ongoing way). Thus the acts of relating are not mere means to a separable end—possibly inefficient means, in fact, to achieving the state of being related. Rather, they constitute what it *is* to be related. In such cases, the very idea of a technological fix coming along that would render the activities of relating irrelevant—except as part of a game we might choose to play—seems to be misguided. Technology cannot *improve* on the activity, if the activity itself is what has intrinsic value (and not as a

means to something distinct). Relating to one another is not a means to something else: it is the goal itself.

This puts me in mind of the tradition that even in the messianic era we Jews will continue to observe Passover.[20] I am uncertain about the sages' own reasons for thinking this, but speaking personally the reason I find the idea attractive is this: at the Passover celebration we sit around the dinner table, eating a fine meal, telling the story of the Exodus from Egypt, and arguing about its meaning. We engage in a discussion about life and death, freedom and responsibility, slavery and salvation. We talk about what is important.

This will be worth doing even when the Messiah comes, even when we enter Utopia. This too is part of the ideal of human existence: people talking with one another.

[20] See, for example, the Babylonian Talmud, tractate Berakhot, 12b.

Bibliography of Robert Merrihew Adams

Books

1987 *The Virtue of Faith and Other Essays in Philosophical Theology* (New York: Oxford University Press).

1994 *Leibniz: Determinist, Theist, Idealist* (New York: Oxford University Press).

1999 *Finite and Infinite Goods: A Framework for Ethics* (New York: Oxford University Press).

2006 *A Theory of Virtue: Excellence in Being for the Good* (Oxford: Clarendon Press).

Research Articles, Discussions, and Occasional Publications

1963 'Anticipation and Consummation', *Theology Today*, 20: 196–211.

1967 'Trust in God', *The Pulpit*, 38: 21–3.

1971 'The Logical Structure of Anselm's Arguments', *Philosophical Review*, 80: 28–54.

——— 'Has it Been Proved that All Real Existence is Contingent?', *American Philosophical Quarterly*, 8: 284–91.

1972 'Must God Create the Best?', *Philosophical Review*, 81: 317–32.

1973 'Berkeley's "Notion" of Spiritual Substance', *Archiv für Geschichte der Philosophie*, 55: 47–69.

——— 'Middle Knowledge' (an abstract), *Journal of Philosophy*, 70: 552–4.

——— 'A Modified Divine Command Theory of Ethical Wrongness', in Gene Outka and John P. Reeder, Jr (eds.), *Religion and Morality* (New York: Doubleday Anchor), 318–47.

1974 'Theories of Actuality', *Noûs*, 8: 211–31.

1975 'Where Do Our Ideas Come From?—Descartes vs. Locke', in Stephen P. Stich (ed.), *Innate Ideas* (Berkeley, Calif.: University of California Press), 71–87.

1976 'Kierkegaard's Arguments against Objective Reasoning in Religion', *Monist*, 60: 228–43.

——— 'Motive Utilitarianism', *Journal of Philosophy*, 73: 467–81.

1977 'Critical Study: *The Nature of Necessity* (A. Plantinga)', *Noûs*, 11: 175–91.

——— 'Leibniz's Theories of Contingency', *Rice University Studies*, 63: 1–41.

——— 'Middle Knowledge and the Problem of Evil', *American Philosophical Quarterly*, 14: 109–17.

1979 'Autonomy and Theological Ethics', *Religious Studies*, 15: 191–4.
——— 'Benevolence and Pleasure', *Reformed Journal*, 29: 13–14.
——— 'Divine Command Metaethics Modified Again', *Journal of Religious Ethics*, 7: 66–79.
——— 'Existence, Self-Interest, and the Problem of Evil', *Noûs*, 13: 53–65.
——— 'Moral Arguments for Theistic Belief', in C. F. Delaney (ed.), *Rationality and Religious Belief* (Notre Dame, Ind.: University of Notre Dame Press), 116–40.
——— 'Primitive Thisness and Primitive Identity', *Journal of Philosophy*, 76: 5–26.
1980 'The Annointing at Bethany', *Princeton Seminary Bulletin* (1980): 51–3.
——— 'Pure Love', *Journal of Religious Ethics*, 8: 83–99.
1981 'Actualism and Thisness', *Synthese*, 49: 3–41.
1983 'Phenomenalism and Corporeal Substance in Leibniz', *Midwest Studies in Philosophy*, 8: 217–57.
——— 'Divine Necessity', *Journal of Philosophy*, 80: 741–52.
——— 'Knowledge and Self: A Correspondence between Robert M. Adams and Hector-Neri Castaneda', in James E. Tomberlin (ed.), *Agent, Language and the Structure of the World: Essays Presented to Hector-Neri Castaneda, with His Replies* (Indianapolis, Ind.: Hackett Publishing Company), 293–309.
1984 'Saints', *Journal of Philosophy*, 81: 392–401.
——— 'The Virtue of Faith', *Faith and Philosophy*, 1: 3–15.
1985 'Involuntary Sins', *Philosophical Review*, 94: 3–31.
——— 'Plantinga on the Problem of Evil', in James Tomberlin and Peter van Inwagen (eds.), *Alvin Plantinga* (Dordrecht: Reidel), 225–55.
——— 'Predication, Truth, and Trans-World Identity in Leibniz', in James Bogen and James E. McGuire (eds.), *How Things Are: Studies in Predication and the History and Philosophy of Science* (Dordrecht: Reidel), 235–83.
1986 'The Problem of Total Devotion', in Robert Audi and William Wainwright (eds.), *Rationality, Religious Belief, and Moral Commitment* (Ithaca, NY: Cornell University Press), 169–94.
——— 'Time and Thisness', *Midwest Studies in Philosophy*, 11: 315–29.
1987 'Berkeley and Epistemology', in Ernest Sosa (ed.), *Essays on the Philosophy of George Berkeley* (Dordrecht: Reidel), 143–61.
——— 'Divine Commands and the Social Nature of Obligation', *Faith and Philosophy*, 4: 262–75.
——— 'Flavors, Colors, and God', in Adams, *The Virtue of Faith*, 243–62.
——— 'The Leap of Faith', in Adams, *The Virtue of Faith*, 42–7.
——— 'Vocation', *Faith and Philosophy*, 4: 448–62.
1988 'Christian Liberty', in Thomas V. Morris (ed.), *Philosophy and the Christian Faith* (Notre Dame, Ind.: University of Notre Dame Press), 151–71.

——— 'Common Projects and Moral Virtue', *Midwest Studies in Philosophy*, 13: 297–307.
——— 'Presumption and the Necessary Existence of God', *Noûs*, 22: 19–32.
1989 'Reply: Cobb on Ultimate Reality', in Linda J. Tessier (ed.), *Concepts of the Ultimate* (London: Macmillan), 52–4.
——— 'Reply to Kvanvig', *Philosophy and Phenomenological Research*, 50: 299–301.
——— 'Should Ethics Be More Impersonal? A Critical Notice of Derek Parfit, *Reasons and Persons*', *Philosophical Review*, 98: 439–84.
1990 'The Knight of Faith', *Faith and Philosophy*, 7: 383–95.
1991 'An Anti-Molinist Argument', *Philosophical Perspectives*, 5: 343–53.
1992 'Idolatry and the Invisibility of God', in Shlomo Biderman and Ben Ami Scharfstein (eds.), *Interpretation in Religion* (Leiden: E. J. Brill), 39–52.
——— 'Miracles, Laws of Nature, and Causation', *Proceedings of the Aristotelian Society*, suppl. vol., 66: 207–24.
——— 'Platonism and Naturalism: Options for a Theocentric Ethics', in Joseph Runzo (ed.), *Ethics, Religion, and the Good Society: New Directions in a Pluralistic World* (Louisville, Ky: Westminster John Knox Press), 22–42.
1993 'Form und Materie bei Leibniz: die mittleren Jahre', *Studia Leibnitiana*, 25: 132–52.
——— 'Prospects for a Metaethical Argument for Theism: A Response to Stephen J. Sullivan', *Journal of Religious Ethics*, 21: 313–18.
——— 'Religion after Babel', in Arvind Sharma (ed.), *God, Truth, and Reality: Essays in Honour of John Hick* (London: Macmillan), 62–71.
——— 'Religious Ethics in a Pluralistic Society', in Gene Outka and John P. Reeder, Jr (eds.), *Prospects for a Common Morality* (Princeton, NJ: Princeton University Press), 93–113.
——— 'Truth and Subjectivity', in Eleanore Stump (ed.), *Reasoned Faith: Essays in Philosophical Theology in Honor of Norman Kretzmann* (Ithaca, NY: Cornell University Press), 5–41.
1994 'Leibniz and the Limits of Mechanism', in *Leibniz und Europa*, VI, *Internationaler Leibniz-Kongreß: Vorträge*, I (Hannover: Gottfried-Wilhelm-Leibniz-Gesellschaft), 1–8.
——— 'Leibniz, Gottfried Wilhelm', in Jaegwon Kim and Ernest Sosa (eds.), *A Companion to Metaphysics* (Oxford: Blackwell), 268–71.
——— 'Leibniz's *Examination of the Christian Religion*', *Faith and Philosophy*, 11: 517–46.
——— 'Religious Disagreements and Doxastic Practices', *Philosophy and Phenomenological Research*, 54: 885–90.
——— 'Theodicy and Divine Intervention', in Thomas F. Tracy (ed.), *The God Who Acts: Philosophical and Theological Explorations* (University Park, Penn.: Pennsylvania State University Press), 31–40.

1995 'Moral Faith', *Journal of Philosophy*, 92: 75–95.

——— 'Introductory Note to **1970*' [i.e. to Gödel's 'Ontological Proof'], in Kurt Gödel, *Collected Works*, vol. iii, ed. Solomon Feferman, et al. (New York: Oxford University Press), 388–402.

——— 'Moral Horror and the Sacred', *Journal of Religious Ethics*, 23: 201–24.

——— 'Agape', 'possible worlds', 'theodicy', and 'transcendence', in Robert Audi (ed.), *The Cambridge Dictionary of Philosophy* (Cambridge: Cambridge University Press), 12, 633–4, 794–5, 807–8.

——— 'Qualia', *Faith and Philosophy*, 12: 472–4.

——— 'Analytical Philosophy and Theism: Reflections on Analytical Philosophical Theology', in William J. Wainwright (ed.), *God, Philosophy, and Academic Culture: A Discussion between Scholars in the AAR and the APA* (Atlanta, Ga.: Scholars Press, 1996), 79–87.

1996 'The Concept of a Divine Command', in D. Z. Phillips (ed.), *Religion and Morality* (New York: St Martin's Press), 59–80.

——— 'Philosophy of Religion', in Donald M. Borchert (ed.), *The Encyclopedia of Philosophy, Supplement* (New York: Macmillan), 427–31.

——— 'The Pre-established Harmony and the Philosophy of Mind', in Roger S. Woolhouse (ed.), *Leibniz's 'New System' (1695)*, Lessico intellettuale europeo, 68 (Florence: Leo S. Olschki), 1–13.

——— 'Response to Carriero, Mugnai, and Garber', *Leibniz Society Review*, 6: 107–25 (part of Adams, Symposium: *Leibniz: Determinist, Theist, Idealist*).

——— 'Schleiermacher on Evil', *Faith and Philosophy*, 13: 563–83.

1997 'Atoning Transactions', in Stephen T. Davis (ed.), *Philosophy and Theological Discourse* (London: Macmillan), 98–101.

——— 'Critical Study: Sleigh's *Leibniz & Arnauld: A Commentary on Their Correspondence*', *Noûs*, 31: 266–77.

——— 'Symbolic Value', *Midwest Studies in Philosophy*, 21: 1–15.

——— 'Things in Themselves', *Philosophy and Phenomenological Research*, 57: 801–25.

——— 'Thisness and Time Travel', *Philosophia*, 25: 407–15.

1998 'Self-Love and the Vices of Self-Preference', *Faith and Philosophy*, 15: 500–13.

1999 'Original Sin: A Study in the Interaction of Philosophy and Theology', in Francis J. Ambrosio (ed.), *The Question of Christian Philosophy Today* (New York: Fordham University Press), 80–104 (with questions and answers from conference discussion, pp. 104–10).

2000 'God, Possibility, and Kant', *Faith and Philosophy*, 17: 425–40.

——— 'Leibniz's Conception of Religion', in *Proceedings of the Twentieth World Congress of Philosophy*, 7 (Bowling Green, Ohio: Philosophy Documentation Center), 57–70.

——— 'Reading the Silences, Questioning the Terms: A Response to the Focus on Eighteenth-Century Ethics', *Journal of Religious Ethics*, 28: 281–4.

——— 'Stewardship or Generosity?', in Wallace M. Alston, Jr (ed.), *Theology in the Service of the Church: Essays in Honor of Thomas W. Gillespie* (Grand Rapids, Mich.: Eerdmans), 12–18.

2001 'Holy Places', *Princeton Seminary Bulletin*, 22: 11–15.

——— 'Scanlon's Contractualism: Critical Notice of T. M. Scanlon, *What We Owe to Each Other*', *Philosophical Review*, 110: 563–86.

2002 'Précis of *Finite and Infinite Goods*', *Philosophy and Phenomenological Research*, 64: 439–44 (part of a book symposium on Adams, *Finite and Infinite Goods*).

——— 'Responses', *Philosophy and Phenomenological Research*, 64: 475–90 (part of a book symposium on Adams, *Finite and Infinite Goods*).

——— 'Science, Metaphysics, and Reality', in Hans Poser (ed.), vii, *Internationaler Leibniz-Kongreß: Nihil sine ratione*, suppl. vol. (Hannover: G.-W.-Leibniz Gesellschaft, 2002), 50–64.

2003 'Anti-Consequentialism and the Transcendence of the Good' (a response to Richard Boyd), *Philosophy and Phenomenological Research*, 67: 114–32.

——— 'The Silence of God in the Thought of Martin Buber', *Philosophia*, 30: 51–68.

2004 'Voluntarism and the Shape of a History', *Utilitas*, 16: 124–32.

2005 'Faith and Religious Knowledge', in Jacqueline Mariña (ed.), *The Cambridge Companion to Friedrich Schleiermacher* (Cambridge: Cambridge University Press), 35–51.

——— 'Human Nature, Christian Vocation, and the Sexes', in Nicholas Coulton (ed.), *The Bible, the Church and Homosexuality* (London: Darton, Longman and Todd Ltd.), 100–13.

——— 'Moral Necessity', in Donald Rutherford and J. A. Cover (eds.), *Leibniz: Nature and Freedom* (New York: Oxford University Press), 181–93.

2006 'How Can I Give You Up, O Ephraim?', *Theology Today*, 63: 88–93.

——— 'Love and the Problem of Evil', *Philosophia*, 34: 243–51.

2007 'Idealism Vindicated', in Dean Zimmerman and Peter van Inwagen (eds.), *Persons: Human and Divine* (Oxford: Clarendon Press), 35–54.

——— 'The Priority of the Perfect in the Philosophical Theology of the Continental Rationalists', in Michael Ayers (ed.), *Rationalism, Platonism and God: A Symposium on Early Modern Philosophy*, Proceedings of the British Academy, 149 (Oxford: Oxford University Press, 2007), 91–116.

Books and Texts Edited, Translated, or Introduced

1975 'The Locke-Leibniz Debate' (edited), in Stephen P. Stich (ed.), *Innate Ideas* (Berkeley, Calif.: University of California Press), 37–67.

1979 George Berkeley, *Three Dialogues between Hylas and Philonous* (edited, with introduction (16 pp.) and other aids to study, by Robert Merrihew Adams) (Indianapolis, Ind.: Hackett).

1990 *The Problem of Evil* (edited, with introduction (24 pp.), by Marilyn McCord Adams and Robert Merrihew Adams) (Oxford: Oxford University Press).

1995 'Appendix B: Texts Relating to the Ontological Proof' (translated), in Kurt Gödel, *Collected Works*, iii, ed. Solomon Feferman, et al. (New York: Oxford University Press), 429–37.

1998 *Integrity and Conscience* (Nomos, 40) (edited, with introduction, by Ian Shapiro and Robert [Merrihew] Adams) (New York: New York University Press).

____ Introduction (pp. vii–xxxii) and other front matter (pp. xxxiii–xxxix) to Immanuel Kant, *Religion within the Boundaries of Mere Reason, and Other Writings*, trans. and ed. Allen Wood and George Di Giovanni (Cambridge: Cambridge University Press).

Books Reviewed

1969 Charles Hartshorne, *A Natural Theology for Our Time*, *Philosophical Review*, 78: 129–31.

____ Robert W. Jenson, *The Knowledge of Things Hoped For: The Sense of Theological Discourse*, *Theology Today*, 26: 358–61.

1970 D. Z. Phillips, *The Concept of Prayer*, *Philosophical Review*, 79: 282–4.

1973 D. Z. Phillips, *Faith and Philosophical Enquiry*, *Journal of the American Academy of Religion*, 41: 439–40.

1977 Michael A. Slote, *Metaphysics and Essence*, *Journal of Philosophy*, 74: 301–8.

1978 William L. Rowe, *The Cosmological Argument*, *Philosophical Review*, 87: 445–50.

1980 Ronald M. Green, *Religious Reason: The Rational and Moral Basis of Religious Belief*, *Religious Studies Review*, 6: 183–8.

1985 Richard Swinburne, *Faith and Reason*, *Noûs*, 19: 626–33.

1986 J. L. Mackie, *The Miracle of Theism*, *Philosophical Review*, 95: 309–16.

1988 Benson Mates, *The Philosophy of Leibniz: Metaphysics and Language*, *Mind*, 97: 299–302.

1990 James W. McClendon, Jr, *Ethics* (*Systematic Theology*, i), *Faith and Philosophy*, 7: 117–23.

1996 Nicholas Jolley (ed.), *The Cambridge Companion to Leibniz*, *Philosophical Review*, 105: 245–8.

1998 Donald Rutherford, *Leibniz and the Rational Order of Nature*, *Philosophical Quarterly*, 48: 264–6.

2000 Maria Rosa Antognazza, *Trinità e Incarnazione: Il rapporto tra filosofia e teologia rivelata nel pensiero di Leibniz*, *Leibniz Review*, 10: 53–9.

2001 Gottfried Wilhelm Leibniz, *Scritti filosofici*, ed. and trans. into Italian by Massimo Mugnai and Enrico Pasini, *Leibniz Review*, 11: 25–8.

2002 J. A. Cover and John O'Leary Hawthorne, *Substance and Individuation in Leibniz*, *Mind*, 111: 851–5.

2006 Linda Zagzebski, *Divine Motivation Theory*, *Philosophy and Phenomenological Research*, 73: 493–7.

This page intentionally left blank

Index