John Rawls: the Path to *A Theory of Justice*

By

Andrius Galisanka

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Political Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Bevir, Chair
Professor Kinch Hoekstra
Professor David Hollinger
Professor Hans Sluga

Spring 2013

Abstract

John Rawls: the Path to *A Theory of Justice*

by

Andrius Galisanka

Doctor of Philosophy in Political Science

University of California, Berkeley

Professor Mark Bevir, Chair

This dissertation is an intellectual biography of American political philosopher John Rawls [1921-2002] from his early years to the publication of his classic work, *A Theory of Justice* [1971]. I focus the historical narrative on Rawls's changing conceptions of philosophy: his ways of raising ethical and political questions and justifying answers to them. I pay particular attention to two aspects of the conception of philosophy found in *A Theory of Justice*: its claim that ethical and political positions are defended by showing that all reasonable persons endorse them in their political judgments, and its aspiration to explicate all of these political judgments in terms of principles of justice.

This conception of philosophy was very influential for Anglophone political thought, contributing to the resurgence of analytic political theory in the 1950s and 1960s. I aim to understand the intellectual origins of this influential philosophical approach and thereby shed light on *A Theory of Justice* and contemporary political thought. Taking this historical approach, I follow the development of Rawls's thought, contextualizing him in contemporary traditions and analyzing his numerous private papers recently deposited in the Harvard University Archives.

I argue that, much to our surprise, Rawls's conception of philosophy originated in logical positivism, the very tradition that is thought to have foreclosed the possibility of political thought in the 1940s. Inspired by logical positivists, Rawls modeled ethics on the "method of science," and, taking ethical judgment as "data," tried to formulate principles, or laws, to explicate them. This analogy between reasoning in ethics and reasoning in science provided Rawls with a conception of objectivity: principles of justice were objective if they explicated the considered political judgments of all reasonable persons. This notion of objectivity made possible reasoned discussion on ethical and political issues and required attention to actual political questions. Yet it also committed Rawls to a contestable view that all reasonable persons agree on a sufficient number of political judgments to yield a conception of justice.

This conception of philosophy changed over the following two decades, but, I argue, it remained positivist. In the early 1950s, Rawls drew on linguistic philosophy's conception of ethical reasoning as a practice, and in the late 1950s he was led on the Wittgensteinian path of considering political questions against the background of seeing morality as a form of life. Nevertheless, the influence of his Harvard colleague W.V.O. Quine in the 1960s brought to light Rawls's positivist conception of

philosophy. Rawls continued to justify political principles by the fact that all reasonable persons endorse them in their political judgments.

My historical narrative contests and supplements the traditional interpretations of Rawls as a Kantian or a theorist in the social contract and rational choice theory traditions. It therefore paints a different picture of 20th century Anglophone political thought. But, as I argue in Epilogue, my narrative also helps to illuminate Rawls's shift to *Political Liberalism*. Doing so, I hope it opens new questions about contemporary attempts to define shared political reasons.

Sincere thanks to

Mark Bevir, Kinch Hoekstra, David Hollinger and Hans Sluga

Harvard University Archives for permission to cite John Rawls's private papers

The University of California, Berkeley, Department of Political Science, Institute of International Studies and the Lithuanian Foundation for their financial support

Table of Contents

1

Introduction: John Rawls and the Positivist Tradition

The Guiding Questions

This dissertation is an intellectual biography of John Rawls [1921-2002] from his early years to the publication of his classic work, *A Theory of Justice* [1971].[1] There are many reasons to study Rawls today. His vast influence on political philosophy stems both from an innovative political vision and a promising philosophical approach. Politically, his defense of inviolable human rights and human equality made important contributions to liberal and democratic thought. His conception of equality and the human person revived Kantianism as a political tradition. In all these respects Rawls's vision played a central role in presenting viable alternatives to the dominant contemporary political tradition of utilitarianism.

Rawls's political vision relied on a novel philosophical framework. It is a common understanding that Rawls's early articles and *A Theory of Justice* played an important part in reviving political philosophy from the desolate landscape in which it was left by logical positivism, and did so by offering a new conception of philosophy.[2] Whereas logical positivism denied the possibility of objective ethical and political judgments, Rawls offered a defensible philosophical approach from which ethical and political judgments can be viewed as right or wrong, reasonable or unreasonable, better or worse. But even today this novel philosophical framework is not well understood. Too often, Rawls is seen as a lone figure, ignoring the contemporary philosophical landscape and looking back to the social contract tradition, Kantianism, and, at his most modern, the emerging field of game theory.[3] Yet these intellectual traditions fail to explain the novelty of Rawls's philosophical approach and consequently misunderstand the character of a new kind of political theory.

I focus on Rawls's philosophical approach as it was expressed in *A Theory of Justice*, and follow his intellectual developments with the aim of explaining how this conception came to be what it is. "Conception of philosophy" is an intentionally broad term. It encompasses two

---

[1] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971). All references are to the original edition unless noted otherwise.

[2] See the introductions to Peter Laslett and W.G. Runciman, eds. *Philosophy, Politics and Society* (Oxford: Basil Blackwell, 1963 [1956]), Peter Laslett and W.G. Runciman, eds. *Philosophy, Politics and Society*, 2nd series (Oxford: Basil Blackwell, 1964 [1962]), Peter Laslett and W.G. Runciman, eds. *Philosophy, Politics and Society, 3rd series* (New York: Barnes & Noble, 1967).

[3] For example, Samuel Freeman claims that "Rawls's research agenda was only mildly influenced by the contemporary discussions in moral and political philosophy" and that "Though raised within the Anglo-American analytic tradition in philosophy, Rawls is mainly responding to problems set forth by the major moral and political philosophers since Hobbes." Samuel Freeman, *Rawls* (New York: Routledge, 2007), 12, 28.

suitably broad questions: What is ethical inquiry about? and What, if anything, makes one ethical judgment or argument better than another? Answers to these questions rely on broader views about human beings, their place in nature and the universe, their ability to make ethical judgments, and the nature of valuable things. Differing on these broader views, rival conceptions of philosophy also differ in their understandings of what makes ethical views objective and what needs to be argued to show them as such.

Although I focus my narrative on Rawls's changing conceptions of philosophy, I also aim to show the implications of these changing conceptions on Rawls's ethical and political views. One's views about the nature of philosophy do not imply any particular political vision. Even so, knowledge of Rawls's philosophical visions allows one to understand aspects of his political positions: why he thought political arguments are rational, as well as what political questions and, in rarer instances, political positions these philosophical frameworks suggested to Rawls. For example, Rawls's focus on the basic structure of society was prompted by linguistic philosophy's notion of human practice, and his view that political judgments are guided by the same several reasons was drawn from logical positivism's conception of empirical theory. Pointing to such connections, my narrative also sheds new light on Rawls's political views.

I pay particular attention to two aspects of Rawls's conception of philosophy: its non-foundational justification and its aim to organize our ethical judgments in terms of a theory, or a list of ordered principles. Rawls rejected foundationalist attempts to defend ethical statements by reference to epistemologically certain statements or experiences. Instead, he argued that ethical argument stops when all reasonable persons agree on an issue in question. In *A Theory of Justice*, Rawls called this stopping point "reflective equilibrium," or the state of affairs in which considered judgments of all reasonable persons converge.[4] As he indicated, any such equilibrium is tentative, as any particular judgment which supports it can be changed. When reason-giving stops, it stops not because we attain certain knowledge, but because for the time being we agree on relevant propositions. Convinced that any philosophical argument must proceed from this tentative agreement, Rawls constructed the "original position," a thought experiment meant to articulate this agreement and draw its logical implications.

Rawls's non-foundationalism was of a specific kind: it was paired with the assumption that all reasonable persons agree on a sufficient number of premises to also agree on a conception of justice. Thus throughout Rawls's early writings there is a sense that reasonable persons have similar conceptual frameworks. Unlike the more radical and historical non-foundationalisms of Thomas Kuhn, Imre Lakatos and Stuart Hampshire, Rawls's approach started from an agreement not within the boundaries of any one tradition, but among all reasonable persons. I organize my historical narrative in such a way that it explains how Rawls's particular type of non-foundationalism came about.

I also focus the historical narrative on the second aspect of Rawls's conception of philosophy: his belief that all ethical judgments of reasonable persons can be explicated in terms of a list of ordered principles. Rawls proposed two such principles: that each person has an equal right to the most extensive system of liberty (the first principle), and that social and economic

[4] Rawls, *A Theory of Justice*, 19-21. Reflective equilibrium is also an equilibrium between reasonable persons' considered judgments and the principles of a theory of justice.

inequalities are to be arranged so as to bring the highest benefit to the least advantaged, with the condition that offices are open to all under conditions of fair equality of opportunity (the second principle).[5] The principles were lexically ordered: the first trumped the second in all situations where the two conflicted.[6] This conception of principles appeared to imply a very straightforward relationship between political knowledge and the practice of politics: guided by a proper conception of justice, the philosopher knew in advance not only the reasons relevant in practical political situations but also their relative weights. In many cases, Rawls thought, this knowledge made ethical judgments mechanical: the philosopher did not rely on intuitive ordering of relevant reasons. In Rawls's mind, the knowledge of relevant reasons and their weights was an advance in clarity and self-knowledge.[7] However, as we will see, Rawls conceded that philosophy could not guide politics in this mechanistic way: his theory was only a "guiding framework designed to focus our moral sensibilities and to put before our intuitive capacities more limited and manageable questions for judgment."[8] In the chapters that follow, I trace the origins of this ambitious aspiration and the reasons for its revision.

I chose this narrative focus for several reasons. First, throughout his life, questions about the nature of ethics and political philosophy were Rawls's central concerns: they posed his main dilemmas and tasks and were responsible for the consequent changes in his worldview. It is significant that, up to the early 1960s, his prime philosophical enemies were not utilitarians but emotivists and intuitionists who defended rival philosophical visions. This focus on philosophy is particularly pronounced from the mid-1940s to the late-1950s, when specifically political questions were peripheral to the main philosophical concerns. Thus, short of writing a comprehensive multivolume biography, Rawls's intellectual history can be most faithfully told by focusing it on his changing understandings of philosophy. Second, a great part of the appeal and influence of *A Theory of Justice* came precisely from Rawls's conception of philosophy. To combat emotivism and utilitarianism, Anglophone philosophy in the 1950s lacked not political visions but a philosophical alternative: a way to show that political judgments can be objective. Rawls's non-foundational justification, summarized in the concept of reflective equilibrium, provided this alternative. And third, emphasis on conceptions of philosophy helps explain the criticisms that the argument of *A Theory of Justice* received as well as reasons for which it was reformulated in *Political Liberalism* [1993].[9] Rawls's argument was criticized on many grounds, including its overly egalitarian political vision, but his change of mind in *Political Liberalism* is best explained by the realization that the conceptual frameworks of reasonable persons are not sufficiently similar. In short, emphasis on Rawls's conceptions of philosophy helps us understand his key intellectual developments from 1942 to 1971, and, as I argue in the Epilogue, to 2002.

[5] Rawls, *A Theory of Justice*, 302.
[6] Rawls, *A Theory of Justice*, 42-43.
[7] Rawls, *A Theory of Justice*, 44-45.
[8] Rawls, *A Theory of Justice*, 53.
[9] John Rawls, *Political Liberalism* (New York: Columbia University Press, 2005 [1993]).

The Argument

I argue that Rawls's conception of philosophy developed in the positivist tradition and, despite significant changes due to the influence of linguistic philosophers such as Ludwig Wittgenstein, it remained positivist in 1971.[10] Rawls was an innovator, but an innovator in the positivist tradition, drawing on its themes to formulate his conception of philosophy and making further innovation against this inherited background. In the initial years, Rawls's positivism was characterized by three key features: the analogy between ethical and scientific inquiry, non-foundationalism and limited meaning holism. By the time it reached *A Theory of Justice*, this positivism differed from its predecessor in two significant ways: the analogy between scientific and ethical inquiries was no longer present and the assumption that all reasonable persons have the same conceptual framework was replaced by the belief that the views of all reasonable persons share a family likeness but do not overlap on every single relevant issue. Yet Rawls's belief that reasonable persons' conceptual frameworks would overlap in significant ways shows that, even in 1971, he belonged to the positivist tradition.

The path to *A Theory of Justice* was long and winding. I start describing it with the analysis of Rawls's undergraduate thesis, *A Brief Inquiry into the Meaning of Sin and Faith* [1943]. When the thesis was first published in 2009, it came as a surprise to many that the author of *A Theory of Justice* – a book that contains no religious terms – started his intellectual career as a religious man.[11] Commentators soon dispelled this surprise by pointing out respects in which the political project of Rawls's classic work is continuous with that of his undergraduate thesis.[12] In a similar way, I describe the character of Rawls's religious conception of philosophy in ways that make sense of his later turn to positivism. This turn is intriguing because positivists, and especially the logical positivists on which Rawls drew, were characteristically dismissive of religious claims, calling them nonsensical. However, my historical narrative makes sense of this Protestant turn to positivism, of which Rawls was a part. As I show, there were in fact many similarities in the philosophical approaches of biblical essentialism and positivism. I draw attention to the fact that already in his undergraduate years Rawls thought of philosophy as analysis of Christian experience, took this Christian experience as the ultimate reason in ethical arguments, and expected that all persons – Christians and non-Christians alike – would share these experiences.

Rawls's emphasis on experience was characteristic of liberal Protestantism. Arising in Germany at the end of the 18th century, liberal Protestantism rejected the Bible as the ultimate authority and replaced it with shared Christian experience. Equipped with this understanding,

---

[10] By "linguistic philosophy," I do not mean the approach that started with Gottlob Frege and is characterized by its belief that an account of thought is best attained through a philosophical account of language. For this narrative, see Michael Dummett, *Origins of Analytic Philosophy* (Cambridge, MA: Harvard University Press, 1993). Rather, I have in mind the intellectual tradition that takes shape in the 1940s and 1950s in Cambridge and Oxford and draws inspiration primarily from Ludwig Wittgenstein's idea that language is a practice governed by rules.

[11] John Rawls, *A Brief Inquiry into the Meaning of Sin and Faith*, edited by Thomas Nagel (Cambridge, MA: Harvard University Press, 2009).

[12] Eric Gregory, "Before the Original Position: The Neo-Orthodox Theology of the Young John Rawls," *Journal of Religious Ethics* 35 (2007): 179-206; Robert M. Adams, "The Theological Ethics of the Young Rawls and Its Background," in Rawls, *Meaning of Sin and Faith*, 24-101; David A. Reidy, "Rawls's Religion and Justice as Fairness," *History of Political Thought* 31 (2010): 309-343.

biblical essentialists such as Adolf von Harnack set out to discover this shared experience. Their expectation was that, once all historical dross was removed, they would discover a commonality in all dogmas of Christianity: the "essence of Christianity."

This movement shaped Rawls's thought through his teachers at Princeton. Rawls started treating the shared Christian experience as the source of ultimate appeal in ethical arguments, understood this experience as an intimate contact with God, thought that this experience is shared by Christians and non-Christians alike, and saw it as his task to provide a conceptual framework to explain it. Rawls soon abandoned his wider web of religious beliefs, as experiences in the Army during the Second World War prompted him to question his understanding of God. Unable to find a satisfying answer, Rawls abandoned his belief in God and by 1946 had dropped the main Christian concepts from his framework. Yet, as I show, he did not abandon all liberal Protestant themes. His thinking continued to rely on experience as the ultimate ground of justification, the expectation of finding a shared "essence" in this experience, and a conception of philosophy as an analysis of this shared experience. I conjecture that these persisting biblical essentialist themes made Rawls open to logical positivism when he joined Princeton as a graduate student in 1946.

In Chapter Three, I reveal the origins of Rawls's initial secular philosophical framework in logical positivism. It may be surprising that logical positivism was the beginning of Rawls's secular thought also because this tradition is typically thought to have "killed" moral and political philosophy by the 1940s. However, as I show, positivism was a broader and more diverse tradition than it is typically thought. The movement's two main positions in ethics were indeed not fruitful: logical positivists either attempted to reduce ethical statements to empirical statements, or argued that, not reducible to empirical statements, ethical statements were meaningless and thus nonsensical. But the mid-1940s saw the development of an interesting position at the fringes of the movement – scientism. Defended by Curt John Ducasse, this position did not rely on the analytic-synthetic dichotomy; instead, it applied the logical positivist conception of scientific inquiry to ethics. If the latter could be modeled on the former, Ducasse conjectured, then the objectivity which we attribute to scientific statements could also be attributed to ethical statements.

The logical positivist conception of scientific inquiry would form the kernel of Rawls's conception of philosophy. It would set his main questions and dilemmas, and, in spite of the modifications he would make to it during the following twenty-five years, it would still form the backbone of his argument in *A Theory of Justice*. The positivist conception of philosophy consisted of three main commitments: the claim that ethical inquiry is an empirical inquiry aimed at explicating considered ethical judgments of reasonable persons, the argument that this explication was non-foundational, and the assumption that ethical judgments are in some sense epistemologically basic. Like Popper and the entire logical positivist tradition, Rawls believed that all reasonable persons would agree in making identical basic judgments if only they were placed in the right circumstances.

Initially, this positivist conception of philosophy was promising: by 1950, when Rawls filed his dissertation, it had led him to elaborate the notions of "rational judgments" and "reasonable men." Yet, despite this fruitfulness, positivism led Rawls to a dilemma when he tried to explain why all reasonable persons would agree in their judgments. The positivist analogy

between ethics and science pushed him to a realist argument: to explain the agreement, Rawls posited an "objective factor residing in the inspected [ethical] situation."[13] As Rawls was critical of any moral realist position, this implication of positivism was problematic.

Rawls changed his views over the next twenty-five years, and the first of these changes came about after encountering critics of logical positivism. Ludwig Wittgenstein in Cambridge and ordinary language philosophers in Oxford had been contesting logical positivism since the 1940s, and their students, including Stephen Toulmin and Stuart Hampshire, soon drew on these new ways of thinking to develop their views in political philosophy. If Rawls drew on the logical positivist position to introduce new accounts of objectivity in ethics and political philosophy, thinkers in Cambridge and Oxford – linguistic philosophers – did so appealing to rival ways of thinking.

Rawls spent a year at Christ Church, Oxford in 1952-53, and, as I show in Chapter Four, this encounter led to a change in his conception of philosophy. Linguistic philosophy influenced Rawls's intellectual development in two ways. First, it provided him with the broad conceptual framework to explain the agreement of reasonable persons without appealing to moral realism. Linguistic philosophers understood language and reasoning as practices governed by socially accepted rules. Some, notably Stephen Toulmin, saw ethical reasoning as a practice governed by the goal of adjudicating the conflicting desires of different persons.[14] As a practice, Toulmin argued, ethical reasoning had to be governed by rules; philosophers only had to uncover them. Rawls seized on this understanding of ethical reasoning as a practice, positing a hypothesis that ethical reasoning is indeed governed by rules and setting out to discover them. This understanding of ethics as a human activity governed by a goal allowed Rawls to escape the positivist push to moral realism. He now argued that, since the goal of ethical reasoning was to adjudicate between competing interests, once this goal was achieved, no further reasons had to be given.

Linguistic philosophy also led Rawls to accept that all reasonable persons would agree on such rules only in an overlapping way. It was a common theme among linguistic philosophers that the meaning of a word varies from context to context and that, if forced to define the meaning of the word generally, outside of any specific context, we could at best give several meanings with overlapping similarities or what Wittgenstein called "family resemblances."[15] Rawls agreed with this argument and extended it to the agreement of reasonable persons, thinking that, although reasonable persons would share conceptual frameworks, these frameworks would nonetheless overlap in partial ways.

The concepts of linguistic philosophy provided Rawls with the tools to examine the connections between ethical and political thought and human emotional life. This part of Rawls's intellectual history is the least known, as it lies hidden in his seminar notes and as it fails to break through Jean Piaget's influence in *A Theory of Justice*. Yet, together with Philippa Foot and John

[13] John Rawls, 'A Study in The Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral Worth of Character', Ph.D. diss., Princeton University (1950), 47-8.
[14] Stephen Toulmin, *An Examination of the Place of Reason in Ethics* (Cambridge: Cambridge University Press, 1950), 137.
[15] Ludwig Wittgenstein, *Philosophical Investigations*, 3rd ed., trans. G.E.M. Anscombe (Upper Saddle River, NJ: Prentice Hall, 1958), §§65-67.

N. Findlay, Rawls was one of the first thinkers to engage in these types of exploration and among the first to criticize emotivism and, by extension, logical positivism in this way. These arguments mark the moment at which the philosophical tide began to turn against emotivism in Anglo-American philosophy. Emotivism had claimed that ethical and political reasons are connected to emotions, but that they are connected in a contingent way. According to them, any reason and any moral principle could in principle arouse human emotion. Wittgensteinian philosophers denied this. They thought that human emotions are logically connected to certain moral concepts, and they thought that only certain moral principles could arouse moral emotions.

In Chapter Five, I focus on Rawls's Wittgensteinian investigations in moral psychology, which he introduced in his seminars at Cornell and Harvard in 1958, 1960 and 1962. These arguments in moral psychology helped Rawls define his naturalist position in ethics: moral reasons were extensions of natural feelings that all human beings were expected to develop given normal circumstances of social life. Furthermore, these naturalist explorations in moral psychology helped Rawls explain why all reasonable persons would agree in their considered judgments of justice. Rawls now argued that, given the roughly shared background of natural feelings – what he, following Wittgenstein, called a "form of life" – "all moralities resemble one another in their prima facie principles; they have this sort of family likeness. They resemble one another in their principles ….[16] This inversion of Wittgenstein's argument of family resemblance shows that Rawls interpreted linguistic philosophy's themes against his positivist background: if Wittgenstein intended the family resemblance argument to counter the contemporary essentialism in philosophy, Rawls used the argument to claim that a sufficient degree of agreement obtains despite the apparent disagreement in the opinions of reasonable persons.

If in the late 1950s one could have conceivably thought that Rawls's conception of philosophy would become closer to Wittgenstein's own, the explorations in moral psychology came as close to Wittgenstein as Rawls would ever be. The remaining ten years until the publication of *A Theory of Justice* mark the entrenchment of Rawls's positivism, which in turn make intelligible Rawls's appeal to traditions by which we traditionally know him: Kantianism, the social contract and rational choice theories. This positivist entrenchment was a result of Rawls's engagement with the thought of his Harvard colleague W.V.O. Quine, the "greatest logical positivist."[17]

I describe this development in Chapter Six. Still guided by the belief that all reasonable persons agree, Rawls thought that justification proceeds by gathering the "fixed points" in the judgments and beliefs of all reasonable persons and selecting a theory of justice which explicates most of these fixed points. In this insistence that all reasonable persons agree and in the belief that the task of philosophy is an analysis of the shared judgments of reasonable persons, positivism continued to shape Rawls's thought.

---

[16] John Rawls, "Essay V" [1958-1962]. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 1, 3, 1i. Judging from the Wittgensteinian arguments made in "Essay V," it is from the time period 1958-1962. (Rawls only numbered the recto in his handwritten notes. I marked the recto as (i) and the verso as (ii). So, sheet 1 consists of both 1i and 1ii).

[17] Hilary Putnam, *Realism with a Human Face*, ed. James Conant (Cambridge, MA: Harvard University Press, 1990), 268.

In Chapter Seven, I show how all the previous developments help us better understand *A Theory of Justice*. In particular, these developments illuminate why this work contains such grand aims and yet recognizes that these aims cannot be achieved. *A Theory of Justice* presents itself as a theory of our judgments: it not only highlights reasons that are relevant for justice but also ranks these reasons in terms of importance. On this picture, philosophy guides political decisions mechanically. This, of course, is the grand positivist goal: to construct a theory the laws of which predict our empirical observations. This grand hope rests on two pillar assumptions: that scientific observers agree in their empirical judgments and that the causes specified in the laws are sufficient to explain the effects in all instances in which they operate. Yet, as Rawls acknowledges in *A Theory of Justice*, political philosophy cannot guide political practice in this mechanical way – and this recognition is the result of a long history of adapting the positivist ideal to the insights of linguistic philosophy. That is why *A Theory of Justice* presents itself – arguably unfairly – as a theory of our judgments and nonetheless guides these judgments only in a general direction.

This mismatch between the ideals of political philosophy and its limits also contains the fractures due to which the argument of the book would have to be reformulated. As I argue in the Epilogue, Rawls's belief that the conceptual frameworks of reasonable persons overlap sufficiently was defensible only in light of his belief that philosophy guides judgment in a general direction. In short, Rawls could expect the agreement of reasonable persons as long as he did not expect them to always agree in practice – in judgments about actual situations. How much agreement in these judgments was to be expected was left open. As I show in the Epilogue, in the decade following the publication of *A Theory of Justice*, Rawls was forced to admit that the expectation that all reasonable persons are expected to agree was mistaken. The lack of conceptual agreement was a dilemma for Rawls as he tried to reformulate his theory.

Rawls's response drew on his earlier positivist themes. Unable to maintain the assumption that all reasonable persons agree sufficiently in their conceptual frameworks, Rawls now hypothesized that they agree on the parts of their framework that are relevant for political questions. In short, he posited agreement in the political – but not comprehensive – culture. Through this series of transformations Rawls's positivism survives in *Political Liberalism* and continues to shape American political thought.

Rival Interpretations

The story of Rawls's intellectual development is not usually told as a story of positivism. Indeed, Rawls is often associated with the resurgence of the social contract and Kantian traditions in the 1960s: his argument in *A Theory of Justice* is seen as one that stems from these traditions. Similarly, Rawls was often viewed as a rational choice theorist. I want to clarify how my narrative relates to these alternative stories, as it would be wrong to deny that the 20th century has seen the resurgence or origination of these traditions in political philosophy, that Rawls played some part in these developments, and that Rawls's own work exhibits some features characteristic of these traditions.

To some extent, this discrepancy between my argument and the accepted narratives does not indicate disagreement. In using terms like "Kantianism," interpreters often meant to call attention not to Rawls's conception of philosophy but to other aspects of his thought. In such cases, these traditions do not offer a conception of philosophy and do not rival positivism. Thus, I believe that Rawls was both a positivist and a Kantian, and that he used the tools of game theory or rational choice in his arguments. In many cases, then, the novelty of my narrative stems from the new issues I raise, not from disagreement with other interpreters. Yet this distinction between philosophical and political narratives dispels only some disagreements: Rawls has also been interpreted as a Kantian, social contract, and rational choice theorist in conceptions of philosophy. I think that these interpretations are wrong, and showing why that is the case should both dispel misconceptions and permit a more accurate grasp of Rawls's thought.

In the years immediately after the publication of *A Theory of Justice*, Rawls was often interpreted as a rational choice theorist. According to this interpretation, his aim was to show that principles of justice are justifiable from non-ethical premises: that they are acceptable to rational egoists. Thus Robert Wolff thought that Rawls's intention was to use the tools of rational choice theory to "derive substantive principles from premises that, though not purely formal, are not manifestly material either."[18] Rawls played a role in this misinterpretation, describing the theory of justice as "a part, perhaps the most significant part, of the theory of rational choice."[19] In fact, however, the relationship between the rational choice theory and Rawls's argument is the inverse: Rawls used the tools of rational choice theory in his own arguments. As I show, he was impressed by the clarity and decisiveness of the game theory arguments, and, thinking that this decisiveness stems partly from its use of the figures of rational egoists in a situation of choice, he decided to use these figures in his own argument. Unlike the rational choice theorists, however, Rawls did not think that people in the real world are rational egoists. He described the persons in his thought experiment as "mutually self-interested" to model our considered judgment that questions of justice should be decided by reasons relevant to justice and not, for example, by sympathy or pity.[20] Thus, while using the tools of rational choice, Rawls did not accept the tradition's wider philosophical framework and therefore its implications in ethics.

Rawls has also been understood as a Kantian and to have inspired the resurgence of the Kantian tradition in the 1960s. In fact, Rawls himself described his principles of justice as "highly Kantian in nature," noting that "there is a Kantian interpretation of the conception of justice from which the principles derive."[21] Rawls's role in this resurgence of Kantianism is indeed significant, and his students Christine Korsgaard, Onora O'Neill and Barbara Herman are among the leading minds in this new movement.[22] Indeed, as I argue in Chapter Six, in 1965 the

[18] Robert Paul Wolff, *Understanding Rawls: A Reconstruction and Critique of A Theory of Justice* (Princeton: Princeton University Press, 1977), 20.
[19] Rawls, *A Theory of Justice*, 16.
[20] John Rawls, "Essay on Justice. First Draft of *A Theory of Justice*, 1 of 2" (1964). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 17, Folder 2, 75.
[21] Rawls, *A Theory of Justice*, viii and 251, respectively. See especially Ibid., 251-257.
[22] For Rawls's Kantian students, see Christine Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996); Barbara Herman, *The Practice of Moral Judgment* (Cambridge, MA: Harvard University Press, 1993); Onora O'Neill, *Toward Justice and Virtue: A Constructive Account of Practical Reasoning* (Cambridge: Cambridge University Press, 1996). For Rawls's own articles, see John Rawls, "Kantian

premises of Rawls's argument – the description of the considered judgments of reasonable persons – became notably Kantian. Nonetheless, Rawls did not become a Kantian in his conception of philosophy. Detached from its metaphysical framework, contemporary Kantian political philosophy is best described by two key claims. First, it is an attempt to derive ethical conclusions from considerations about what it is to make an ethical judgment, or, more broadly, what it is to take the standpoint of practical reason. Second, these conclusions are viewed as necessary given that taking the practical standpoint is unavoidable.[23]

When the features of Kantianism are specified, it becomes apparent that Rawls was not a Kantian in his conception of philosophy. As is evident even in his earliest writings, Rawls was a non-foundationalist in his philosophy and believed that no judgment or statement is safe from testing and revision. Therefore, he did not believe that a conception of justice could be a necessary implication of making an ethical judgment, partly because the conditions of making a judgment could be described in various ways. But more importantly, Rawls did not think that principles of justice should be seen as implications of taking a practical point of view. He made use of some implications of making a judgment – the conditions of universality and finality – but he did not think that, taken by themselves, these implications would be sufficient to deduce a conception of justice. Indeed, in Chapter Five I show that Rawls criticized Kurt Baier and Richard Hare precisely for trying to derive ethical principles from the conditions of making a judgment alone. Despite Rawls's impact on the development of contemporary Kantianism, Rawls did not defend justice as fairness in a Kantian way.

Most plausibly but nonetheless mistakenly, Rawls has been understood as a social contract theorist, and in particular one of the originators of the contemporary contractarian tradition. Rawls persistently defended this interpretation, describing his theory as an attempt to "generalize and carry to a higher order of abstraction the traditional theory of the social contract."[24] This interpretation rests on Rawls's requirement that a conception of justice be acceptable to all reasonable persons in a well-ordered society. Samuel Freeman defended this "contractarian" interpretation, claiming that "[it] is this general agreement among the members of a well-ordered society that mainly drives the contractarian element in Rawls's view…."[25]

Again, however, Rawls's conception of philosophy in *A Theory of Justice* does not align with that of the social contract tradition, even if this latter is understood in the contemporary contractarian way. The most defensible contemporary contractarian conception of philosophy revolves around the idea of acceptability without agreement: it attempts to forge a conception of justice that is acceptable to all although not held by all. On this view, political philosophy is a practical enterprise aimed at securing feasible agreement. Thomas M. Scanlon's distinction

---

Constructivism in Moral Theory," in John Rawls, *Collected Papers*. Edited by Samuel Freeman (Cambridge, MA: Harvard University Press, 1999), 303-358.

[23] See Robert S. Taylor, *Reconstructing Rawls: the Kantian Foundations of Justice as Fairness* (University Park, PA: the Pennsylvania State University Press, 2011), 234.

[24] Rawls, *A Theory of Justice*, viii.

[25] Samuel Freeman, *Justice and the Social Contract: Essays on Rawlsian Political Philosophy* (Oxford: Oxford University Press, 2007), 4.

between a position that is one's own and a position that cannot be reasonably rejected, is a good example of contractarianism's main idea.[26]

Yet this core contractarian expectation that political philosophy is a practical activity involving compromises plays no role in Rawls's argument. There is no need for a compromise along the lines described above. This can be seen from a contrast between *A Theory of Justice* and Rawls's later argument in *The Law of Peoples*.[27] In this later book, in which Rawls rejected the key positivist assumption that all reasonable persons would come to have sufficiently similar conceptual structures, he argued that different reasonable people would justify the law of peoples from their own points of view, using different kinds of arguments.[28] The argument from the original position was treated as only one of the possible arguments for the laws of peoples. But this acknowledgement that politics involves acceptance that the best conception of justice might be one that is acceptable to all though not held by all is missing in *A Theory of Justice*.

Intellectual Histories and Grand Narratives

Narrating Rawls's intellectual history, I follow the Cambridge school approach to intellectual history, which explains thinkers by placing them in relevant contemporary intellectual traditions and identifying their dilemmas and responses.[29] Cambridge historians start from several assumptions. First, they think that our beliefs are always formed in particular historical contexts and are consequently informed by past intellectual commitments. Consequently, they view each human being as a thinker of his or her own, interpreting what is said against the background of beliefs already held. As initial beliefs and later experiences differ, often even the most earnest attempts to understand lead to differences in belief. Taking this insight into human understanding seriously, Cambridge school historians start with individuals and draw the boundaries of larger intellectual movements – traditions of thought – around them. As some thinkers share most of the characteristic beliefs of the tradition but differ in other respects, these boundaries will not be very precise, and, moreover, not all of its thinkers will necessarily have any one commitment in common. Building on Ludwig Wittgenstein's ideas, the Cambridge school assumes that, for a tradition to count as a tradition, it is sufficient that its members' beliefs share "family likenesses" despite not holding any one commitment or question in common.[30] To emphasize the fact that intellectual traditions are drawn around individuals, historians have called these traditions "aggregate concepts."[31] Lastly, Cambridge historians think that the most accurate way to explain change is to outline the thinker's main commitments, call attention to 'dilemmas' which stem from inconsistencies between some of these commitments, and, finally, describe beliefs that result from the solution of these dilemmas. Narrating Rawls's

[26] Thomas M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998), 189-247.
[27] John Rawls, *The Law of Peoples* (Cambridge, MA: Harvard University Press, 2001).
[28] Rawls, *Law of Peoples*, 30-34.
[29] For methodological discussions, see Mark Bevir, *Logic of the History of Ideas* (Cambridge: Cambridge University Press, 1999) and James Tully, ed, *Meaning and Context: Quentin Skinner and his Critics* (Princeton: Princeton University Press, 1988).
[30] Wittgenstein, *Philosophical Investigations*, §§ 66-67.
[31] Mark Bevir, "Political Studies as Narrative and Science, 1880-2000," *Political Studies* 54 (2006): 583-606.

intellectual development, I will place him in the intellectual traditions that are most relevant in explaining his main commitments, resulting dilemmas, and consequent changes in his beliefs.

At its best, the Cambridge approach to history combines knowledge of the relevant intellectual currents with extensive analysis of the thinker's own writings. In Rawls's case, this has recently become possible with the opening of his private papers and notes at the Harvard University Archives. Consisting of thousands of pages of unpublished essays and lecture notes, this collection is an invaluable resource to understanding Rawls's intellectual development. Without these private notes, this dissertation would not have been possible. These papers help us open questions about Rawls's thought, his place in 20th century political philosophy and his legacy.

Situating Rawls in relevant contemporary and past intellectual traditions is helpful in understanding not only Rawls's thought, but also the broader developments of 20th century political philosophy and the humanities more broadly. I argue that 20th century political philosophy is also a story of the developing positivist tradition. Seeing Rawls in this light, we can relate his intellectual history to the history of the post-analytic turn in Anglophone philosophy. Scholars have argued that early analytic philosophy can be best described as modernist because of its atomistic and foundational features.[32] They have included G.E. Moore, Bertrand Russell, early Wittgenstein, and early logical positivists under this broad tradition of modernism. However, we can reasonably expand the definition of atomism to include non-foundational approaches which do not draw radical conclusions from meaning holism: this would allow us to include the later logical positivist tradition, and therefore Rawls, into the story of analytic philosophy. Both foundational and non-foundational modernism, when faced with the objections of later Wittgesntein and Quine, were forced to modify atomism and to adopt more historical themes.[33] Although Rawls expressly rejected atomism, he nonetheless drew very limited implications from meaning holism. For this reason, his intellectual history is in line with the broader history of modernism: it contains the same protagonists, involves similar developments, and ends with a turn to a more contextual and more historical inquiry. This broader turn of modernism to the more historical modes of reasoning is intriguing, as it is telling of the shared shortcomings of the movement. My dissertation helps us understand why Rawls, one of the clearest modernist thinkers, and one who started from the most defensible modernist position, was nonetheless also forced to turn to historical themes. Combined with histories of modernism in other disciplines, I hope it helps us make better sense of the 20th century intellectual landscape.

[32] John Skorupski, "The Legacy of Modernism," *Proceedings of the Aristotelian Society* 91 (1990): 1-19.
[33] Richard Rorty, "How Many Grains Make a Heap?" *London Review of Books* 27:2 (20 Jan., 2005); Richard Rorty, "Introduction: Metaphilosophical Difficulties of Linguistic Philosophy" in Richard Rorty, ed. *The Linguistic Turn: Essays in Philosophical Essays* (The University of Chicago Press: Chicago, 1992 [1967]), 1-39.

2

The Liberal Protestant Beginnings

Introduction

John Bordley Rawls was born on February 21, 1921 in Baltimore, Maryland to what he called a conventionally religious family. His father, William Lee Rawls, was a Southern Methodist, but he frequented the Episcopalian Church, the congregational home of his wife and John's mother, Anna Abell Stump. Like his older brother William, Rawls attended the Episcopalian Kent School from 1935 to 1939, when he started his undergraduate education at Princeton University. Throughout his early life Rawls was only conventionally religious, but this changed at Princeton. There, he developed an interest in theological questions, which resulted in an undergraduate thesis, *A Brief Inquiry into the Meaning of Sin and Faith*.

These facts about Rawls's religious life were largely unknown to academia until 2009, when his undergraduate thesis was published. To many, Rawls's religious background initially came as a surprise: Rawls's political and ethical visions in the early articles and *A Theory of Justice* do not use any religious concepts, and even the topic of religion does not occupy a central place. A large gap seemed to lurk between Rawls's religious thought and his later, secular political writings. Yet in other ways, Rawls's life story is a familiar example of the continuing secularization of 20th century Protestant America. It also exemplifies the continuation of Protestant themes in the secular American landscape.

American Protestant theology in the 1930s and 1940s was varied and lively. Detailing the currents that shaped Rawls's early commitments is therefore a difficult task. His thesis reflects these turbulent inter-war decades, defined by American Protestantism's attempt to establish the basis of its universalism. In an important way, the thesis shows the decline of the liberal Protestant attempt to base Christian universalism on the shared experiences of the Christian community and the rise of neo-orthodoxy, which turned American Protestant universalism back to supernatural revelation. Rawls's thesis draws on liberal Protestantism and its reliance on the shared Christian experiences; but it is a liberal Protestantism which by 1940s had become dehistoricized and ever-more empirical in its attempt to avoid the neo-orthodox objections. Rawls's theological commitments and his later secular arguments are part of this 20th century shift away from historicism.

The existing literature on Rawls tells a different story of his thesis and Protestant theology: it portrays Rawls as a neo-orthodox thinker. Eric Gregory and Robert Adams emphasize neo-orthodoxy's influence on Rawls's view of God as a person and on his claim that

full knowledge of God is possible only through his self-revelation.[1] David Reidy calls attention to Rawls's statement that the Bible is "the last word in matters of religion," a statement he interprets as a typical neo-orthodox theme.[2] However, I think that, once these statements are put into their proper context, they no longer appear neo-orthodox. Rawls did not actually believe that the Bible is the standard of truth, only that it was a correct record of Christian experiences. His thought was indeed influenced by aspects of neo-orthodoxy, but these aspects are not sufficient to explain the main goals of Rawls's thesis.

To explain these main goals, we need to tell a different story: one that centers on an important current in liberal Protestantism – biblical historicism. Characterized by its reliance on Christian experience as the ultimate source of appeal and its belief that this experience is shared by all Christians, biblical historicism shaped Rawls's main commitments and the two main goals of the thesis: to develop a conceptual framework that analyzes shared Christian experience correctly and to show other such analyses mistaken.

To make this argument, I will first delineate American Protestant theology by the 1940s. When Rawls entered Princeton in 1939, neo-orthodoxy was at the height of its influence and biblical historicism had waned as a movement. In particular, biblical historicism had lost its historicist edge. What remained was its emphasis on the agreement of all Christian persons – without the claim that this agreement can be shown only after analyzing the historical development of the Christian tradition. I show this eclipse of historicism and the rise of neo-orthodoxy first in Germany and Switzerland, where the movements originated, and then in the United States, to which these movements traveled in the 19th and 20th centuries. Turning to *Meaning of Sin and Faith* thereafter, I will show how biblical historicist themes help us understand some of the dilemmas Rawls faced as a soldier during the Second World War, as well as some of the reasons for which he eventually abandoned this early religious framework. Biblical historicism will also help us explain Rawls's turn to positivism and illuminate core aspects of his later thought, in particular his conception of philosophy as analysis of our judgments, the underlying belief that these judgments, if analyzed correctly, will reveal commonly shared principles, as well as reliance on these judgments as ultimate justification.

The Rise and Fall of Liberal Protestant Historicism in Europe

Rawls's undergraduate thesis is a reflection of the rise and fall of historical liberal Protestantism. It starts with the claim that the theologian's task is analysis of experience and proceeds to investigate this experience in an ahistorical way: not distinguishing between experiences of different Protestants, Christians, or, indeed, between religious and secular thinkers.[3] All are thought to have these experiences. The thesis grounds the universalism of its theory in these shared experiences: the theory is said to be correct because it is thought to explain these shared experiences.

[1] Gregory, "Before the Original Position," 185-8; Adams, "Theological Ethics of the Young Rawls," 25-32.
[2] Reidy, "Rawls's Religion," 315-16; Rawls, *Meaning of Sin and Faith*, 254.
[3] Rawls, *Meaning of Sin and Faith*, 110.

While this nearly empiricist conception of theology was not usual in the context of American Protestantism, it nonetheless drew on the familiar liberal Protestant themes, and in particular the dehistoricized biblical historicist and the biblical essentialist traditions. To understand the main commitments and goals of Rawls's undergraduate thesis, we need to understand the demise of this liberal Protestant historicism and its more empiricist or positivist varieties in the 1940s's America.

Liberal Protestantism as a tradition of thought is best defined by its claim that the truth of the Christian message rests not on external authority but on its ability to explain the experiences of the Christian community.[4] Liberal Protestantism contains many currents; in the 20th century American Protestant theology alone one discerns pragmatic, personalist, and what is usually called Ritschlian and Harnackian but what I will name biblical historicist currents.[5] These currents give the Christian experiences different philosophical significance, and, consequently, describe them differently. Thus, the label of liberal Protestantism subsumes varying and disagreeing currents of thought even though they all reject orthodox or neo-orthodox conceptions of theology. To explain Rawls's early religious framework, we need to refer to biblical historicism and biblical essentialism, mentioning other liberal Protestant currents only as needed.

Biblical historicism and biblical essentialism arose in the nineteenth and early twentieth centuries in part as a response to the dilemmas raised by the studies of biblical contexts. Revealing that the Bible was written in different times by different people, these studies questioned the Bible's status as a record of a revelation that happened once and was written at once.[6] For this and other reasons, biblical scholars were forced to conclude that the Bible was often historically inaccurate, some, such as W.M.L. de Wette [1780-1849], going as far as to say that "not one of the historical books of the Old Testament has any historical value."[7] Studies of biblical contexts also revealed the foreign nature of Christ's message. Hermann Reimarus [1694-1768] argued that, contrary to the accepted belief, the historical Jesus promised deliverance from the Roman Empire, not a spiritual redemption of the people of Israel.[8] In sum, studies of biblical contexts raised two main dilemmas: the Bible's status as the standard of truth was threatened, and the message of Jesus, insofar as it was understood literally, seemed hardly acceptable.

Biblical historicism and biblical essentialism originated in part as a response to these historical findings and dilemmas. Rejecting the prevalent doctrine of Biblical inerrancy as well as the attendant conception of the Bible as the standard of truth, historicist thinkers were at a loss to justify contemporary expressions of Christianity. Its answers to these dilemmas are best exemplified by the writings of Albrecht Ritschl [1822-1889], a biblical historicist, and Adolf von Harnack [1851-1930], who is best understood as a biblical essentialist. Departing from the

---

[4] See Gary Dorrien, *The Making of American Liberal Theology: Imagining Progressive Religion 1805 – 1900* (Louisville, KY, 2001), xiii, 1.

[5] See Gary Dorrien, *The Making of American Liberal Theology: Idealism, Realism, and Modernity 1900-1950* (Louisville, KY: Westminster John Knox Press, 2003), 21-72, 216-355.

[6] Thomas A. Howard, *Religion and the Rise of Historicism* (Cambridge: Cambridge University Press, 2000), 39-40. Scholars of biblical contexts did not themselves turn to biblical historicism. Some of them concluded that the Bible contained both myths and historical facts; see Howards' account of J.G. Eichhorn and W.M.L. de Wette in Howard, *Religion and the Rise of Historicism*, 36-38.

[7] Quoted in Howard, *Religion and the Rise of Historicism*, 38.

[8] Howard, *Religion and the Rise of Historicism*, 82.

orthodox theme of biblical inerrancy, Ritschl and Harnack argued that the truth of the Christian message rests not on the accuracy of its original revelation as recorded in the Bible but rather on Christianity's ability to accord with the experiences of the Christian community. This reliance on the Christian experience, while not novel in the German Protestant context, was radically new in light of the orthodox conception of theology.[9]

Ritschl's and Harnack's reliance on Christian experience differed and these differences are important for understanding American Protestant theology and Rawls's early thought. Harnack insisted that the Christian community – past and present – has shared experience. This Christian experience, according to him, had a "kernel" or an "essence"; as he wrote, "certain fundamental ideas of the Gospel have never been lost and have defied all attacks."[10] However, as this essence was not apparent in the divergent historical expressions of Christianity, Harnack argued that it was the task of a theologian to study these different expressions and reveal the commonality that hides in them. Ritschl, on the other hand, appealed to neo-Hegelian developmental historicism to claim that Christianity is the end of all religions. Thus for him it was not important if Christianity significantly differed from other religions. Instead, he argued, "in Christianity the tendency of all the [historical religions] finds its perfect consummation"[11] Both Ritschl and Harnack saw the role for history, but Ritschl was importantly historicist where Harnack was not: he thought that the very development of history made a difference, since Christianity in its most defensible form would be found only at the later stages of historical development. Harnack, on the other hand, saw historical differences as dross which hid an essence common to all Christian doctrines. In these different ways, both thinkers defended Christian universalism by referring to Christian experiences.

In their descriptions of the Christian experiences, biblical historicists and essentialists appealed to Friedrich Schleiermacher [1768-1834] and G.W.F. Hegel [1770-1831] to argue that experience can be both non-intellectual and intellectual. For Schleiermacher, who distinguished religious experience from intellectual insight, religious experience consisted in an expressly anti-intellectual "intuition and feeling."[12] For Hegel, on the contrary, religion was rational, aiming to understand the nature of things; consequently, he thought that the truth of religion is to be found in its dogmas, not in religious feeling.[13] Ritschl and Harnack combined these views. Harnack, albeit prioritizing Christian dogmas and teachings, nonetheless insisted that Christianity was

[9] Friedrich Schleiermacher is generally thought to be the first thinker to place the weight of justification on Christian experience. See his *On Religion: Speeches to its cultured despisers*, translated and edited by Richard Crouter (Cambridge, Cambridge University Press, 1996). But Schleiermacher's approach differs from Ritschl's and Harnack's in not being historical, for which reason I do not start the narrative of biblical historicism with him.
[10] Adolf von Harnack, *Outlines of the History of Dogma*, Edwin Knox Mitchell, trans. (Boston, 1957), 7-8 (emphasis original).
[11] Albrecht Ritschl, *The Christian Doctrine of Justification and Reconciliation*, translated by H.R. MacIntosh and A.B. Macaulay (Edinburgh: T. & T. Clark, 1902), 197.
[12] Schleiermacher, *On Religion*, 22. This intuition was expressly anti-intellectual, not meant to provide knowledge of God: while it did intuit a certain object outside of itself, it did not, Schleiermacher wrote, attempt to explain it. Schleiermacher, *On Religion*, 22-25.
[13] See Williams Adams Brown, *The Essence of Christianity* (New York: Charles Scribner's Sons, 1902), 192-6.

more than systems of thought and included experiences that Christ and his followers inspired in other members of the Christian community.[14]

Liberal Protestant thinkers in general differed in their conceptions of Christ and the Bible, but for the most part, they radically humanized Christianity and cast its supernatural aspects away. They viewed Christ as one human being among others: while Schleiermacher still called him the mediator between God and humanity, Harnack thought that Christians distinguished Christ from other human beings not by his likeness to God but by the delivery of his message, which, itself not new, was proclaimed "with exceptional strength and vigor."[15] Similarly, historicists viewed the Bible not as a record of revelation and therefore the standard of truth, but as a historical document intended to bring to light and to call forth the Christian experiences, which, in turn, were expected to "awaken a belief in Jesus Christ's person and mission."[16]

Neo-Hegelian biblical historicism and Harnack's biblical essentialism swept the Protestant landscape. Harnack was particularly influential. However, by the 1920s and 30s these traditions gave way to a new movement, which called itself the "theology of crisis," but which, following American intellectual historians who studied its influence on Rawls, I will call neo-orthodoxy. Neo-orthodoxy originated largely from within biblical historicism and essentialism, as convinced historicists turned against these movements after the start of the First World War. Neo-orthodoxy's dilemmas therefore center on historicism, although answers to these dilemmas and consequent characteristic neo-orthodox themes hardly reveal this origin. To many historicists, including Karl Barth [1886-1968], the First World War made it clear that reliance on Christian experience as the ultimate standard of truth may lead to unacceptable results. He was led to this conclusion after biblical essentialists' public endorsement of Germany's decision to join the war, which they justified by religious experience. Barth's teacher Harnack was among the signatories of this document.[17] Judging this support deeply mistaken, Barth inferred from it that "the exegetical and dogmatic presuppositions [of Biblical essentialists] could not be in order."[18] From that point onward, he wrote, he "could not any longer follow either their ethics and dogmatics or their understanding of the Bible and of history."[19] It took Barth another six years to formulate his dilemmas precisely: he concluded that the signatures were not individual mistakes, but entailments of biblical historicism's and biblical essentialism's emphasis on Christian experience, which, he thought, went all the way back to Schleiermacher.[20]

To solve this dilemma, neo-orthodox thinkers rejected biblical historicist themes: they thought the more direct experiences of God impossible, and the agreement of all Christians as no argument for the truth of Christianity. Instead, these theologians brought back more orthodox themes, such as the supernatural revelation that required God's action to take place and faith to be understood, Christ as the carrier of that revelation, and the Bible as the sole source of the

---

[14] Adolf von Harnack, *What is Christianity?* Thomas Bailey Saunders, trans. (Oxford: Williams and Norgate, 1901), 11.
[15] Schleiermacher, *On Religion*, 121; Harnack, *What is Christianity*, 47-8.
[16] Harnack, *What is Christianity*, 20.
[17] Gary Dorrien, *The Barthian Revolt in Modern Theology* (Louisville, KY: Westminster John Knox Press, 2000), 38.
[18] Barth quoted in Dorrien, *The Barthian Revolt*, 38.
[19] Barth quoted in Dorrien, *The Barthian Revolt*, 38.
[20] Dorrien, *The Barthian Revolt*, 43-4.

complete knowledge of God. So, Barth thought that revelation was not a direct and frequent communion with God but a fact of history: it has happened only once, through Jesus. Consequently, he wrote in a public debate with Harnack, historicists misunderstand the nature of Jesus: "the historical reality of Christ … is not the 'historical Jesus' whom an all too eager historical research had wanted to lay hold of …. Nor is it, as you said, an imagined Christ but rather the *risen* one."[21] Similarly, the changed Barth thought that the Bible was not a "more or less concealed religious possibility of man" but the Word of God.[22] Accordingly, a theologian should talk not of his own personal experiences or those of the Christian community, but of God.[23] Liberal Protestants, attempting to reduce history to the merely human, misunderstood it and emptied the task of theology.[24]

Emil Brunner [1889-1966] was the most influential neo-orthodox thinker outside of the German speaking world and a direct influence on Rawls. Like Barth, he also started out as an essentialist, and in 1928 still acknowledged that historical research is needed to understand God's word in the Bible, agreeing that parts of it are historically inaccurate and that its scriptures are inconsistent.[25] Yet however appreciative of biblical essentialism Brunner continued to be, he now rejected its main themes. While liberal Protestants talked of the experience of dependence on God, Brunner claimed that revelation happened only once, with the apparition of Jesus.[26] Moreover, he thought, this particular historical revelation was entirely God's decision to reveal himself and in no way depended on human beings.[27] Brunner believed that knowledge of any person, including God, was possible only through revelation; thus, he thought that we can know God only because he has revealed himself to us: "God Himself wills to speak to us; we can only perceive that truth in so far as God Himself actually speaks to *us*."[28] He saw Christ as the mediator through which God revealed himself to mankind; thus Christ is not the best of men, but unlike men in essential respects.[29] Like Barth, Brunner thought that this failure to understand revelation and Christ correctly was a result of Biblical historicism's limited starting points: as he put it, "the revelation of Christ does not cease with the processes which the historian can verify – even when he has every possible kind of material at his disposal."[30]

Neo-orthodox thinkers thus rejected key liberal Protestant commitments. This difference can be best summarized by bringing forth the differing conceptions of knowledge of God. As both Barth and Brunner thought that God's self-revelation has happened only once – through the life of Jesus Christ, the mediator – and as, according to them, this revelation was accurately and

---

[21] Karl Barth, "An answer to Professor Adolf von Harnack's open letter," in H. Martin Rumscheidt, *Revelation and Theology: An Analysis of Barth-Harnack Correspondence of 1923* (Cambridge: Cambridge University Press, 1972), 45-6 (emphasis original).
[22] Barth, "Answer to Harnack," 44.
[23] Barth, "Answer to Harnack," 42.
[24] Barth, "Answer to Harnack," 35, 42.
[25] Emile Brunner, *The Theology of Crisis* (New York: Charles Scribner's Sons, 1929), 7, 20; Emil Brunner, *The Mediator: A Study of the Central Doctrine of the Christian Faith*, Olive Wyon, trans. (London: The Lutterworth Press 1934), 34, 170-2.
[26] Brunner, *Mediator*, 30.
[27] Brunner, *Theology of Crisis*, 31-3.
[28] Emil Brunner, *Man in Revolt: A Christian Anthropology*, Olive Wyon, trans. (New York: Charles Scribner's Sons, 1939), 67 (emphasis original).
[29] Brunner, *Mediator*, 50, 39, 64.
[30] Brunner, *Mediator*, 159.

fully recorded in the Bible, they both concluded that one could know God best through the Bible. Brunner argued that there were other ways of knowing God, namely, through what he called "general revelation," or the way in which God discloses himself without intending it. Creation, Brunner thought, was a good example of such unintended self-revelation, and one found in the Bible.[31] Barth thought that this was a mistake, and the two clashed in a famous debate in 1934.[32] Despite this difference, however, both Barth and Brunner agreed that one could not know God through the more personal and direct consciousness of God that the biblical historicists and essentialists emphasized. Christian experiences did not play any role in the neo-orthodox argument. Neo-orthodoxy did not deny that Christians had experiences of repentance or guilt, but it held that these experiences were not a legitimate way of deriving the Christian message. Thus, neo-orthodox thinkers rejected the appeal to Christian experience in principle, and were not interested if it had a common core. This was a radically different conception of God and theology than that of biblical historicism, and in this radical form it would reach the U.S.

Historicism, Essentialism and Neo-orthodox Theology in the United States

American theology in the 1930s and 1940s was thoroughly shaped by this conversation between liberal Protestantism and neo-orthodoxy. Indeed, it is remarkable how influential German theology was on American religious thought, and how similar the rise and demise of historical thought in American liberal Protestantism was to this same process in Europe. Importantly, however, American theology was more receptive of Biblical essentialism. Even at its height, American liberal theology was not historicist. Historical research was used to discover the essence of Christianity, but historical development was not deemed important for the truth of the Christian doctrine.

American reception of Harnack but not of Ritschl is the first step in explaining Rawls's ahistorical conception of philosophy. U.S. American theologians defended Christianity by trying to uncover the common Christian experiences – the essence of Christianity. Despite the initial controversies caused by historical arguments – such as Charles Briggs' [1841-1913] assertions that Moses did not write the Pentateuch or the Book of Job and that, in general, the Bible contained errors – by the early twentieth century biblical essentialism took root in the United States.[33] Concentrating mainly in the Union Theological Seminary in New York and, for a brief period, in the University of Chicago School of Theology, it slowly made the biblical essentialist themes an integral part of the American theological landscape.[34] Thus, theologians like Charles Briggs, William Adams Brown, Henry Churchill King and Shailer Matthews expected to uncover the essence of Christianity by purifying it "from all dross, brushing away the dust of

[31] Brunner, *Mediator*, 32-4; Brunner, *Man in Revolt*, 70-1.
[32] See Brunner's "Nature and Grace" and Barth's "No!" in John Baillie, ed. *Natural Theology* (Eugene, OR: Wipf and Stock Publishers, 2002).
[33] Dorrien, *American Liberal Theology 1805-1900*, 358. Famously, Charles Briggs, accused of heresy, was removed from his ministry in the Presbyterian Church as late as 1893. Ibid., 359-362.
[34] The University of Chicago School of Theology turned pragmatist soon after the arrival of John Dewey to the university's faculty of philosophy.

tradition."[35] They thought that this essence would consist in shared Christian experiences. To emphasize the Schleiermacher-like conception of religious experience, Brown, King and Matthews stressed that Christ's person should be understood as a living spirit revealed in the history of Christian religious experiences.[36] As Brown wrote, "if we are to understand the nature of the Christ of whom we speak, [we must] study of the effects which he has produced in human life. Here our own experience gives us invaluable help…."[37] Biblical essentialists understood Christian experience – and thus revelations of God – broadly. Thus, Brown thought that God's presence is experienced in daily life, but also in the more dramatic experiences of conversion in which "the Christian life begins."[38]

Biblical essentialism in America was also displaced by neo-orthodoxy, which, partly because of Brunner's numerous lectures, had well-known defenders already by the end of the 1920s.[39] The dilemmas that drove American neo-orthodoxy differed from those of the European neo-orthodoxy. Reinhold Niebuhr [1892-1971], for instance, faulted essentialists and historicists for their mistaken belief in the progress of man, not their emphasis on Christian experiences. However, as the American neo-orthodox thinkers appealed to Barth and Brunner to solve their dilemmas, the themes of American neo-orthodoxy did not diverge from those of its European counterpart. Thus, Niebuhr thought that man's evident sinfulness required different conceptions of the Bible, Christ, and revelation. He therefore emphasized the supernatural character of revelation, the inability of reason to fully know God, and Christ as the carrier of God's message and our way of understanding God. Like Brunner, he allowed that we can know God not only from the Bible, but also through contemplating God's creation of the world, which evoked a feeling of being unqualifiedly dependent on God and of "being seen, commanded, judged and known from beyond ourselves."[40] Unlike the European neo-orthodox, Niebuhr did not attack the core biblical historicist and essentialist themes. Consequently, American neo-orthodoxy's historical connection with biblical historicism and essentialism was lost from sight. It is one of the reasons, I think, why commentators on Rawls's *Meaning of Sin and Faith* failed to see biblical essentialist themes in this work.

When Rawls entered Princeton University in 1939, neo-orthodoxy was at the height of its influence and biblical essentialism had already waned as a movement. Yet, much of American theology was not straightforwardly neo-orthodox: influential biblical essentialist works were still being published in English, and many biblical historicist themes persisted, mostly by being incorporated into, or – as was the case with Rawls's teachers – by incorporating themes of other traditions of thought. One among such influential biblical essentialist works was Swedish

---

[35] Charles Briggs quoted in Dorien, *American Liberal Theology 1900-1950*, 351, 340-2.
[36] Dorien, *American Liberal Theology 1900-1950*, 351, 340-42; 66; 196. See also Brown's statement that theology is a "normative science, whose function it is to discriminate that which is essential and permanent in Christian faith from that which is accidental and temporary." William Adams Brown, *Christian Theology in Outline* (New York: Charles Scribner's Sons, 1919 [1906]), 6.
[37] Brown, *Christian Theology in Outline*, 21.
[38] Brown, *Christian Theology in Outline*, 31-32, 408.
[39] In 1919-20, Brunner was awarded a fellowship to teach at the Union Theological Seminary, and in 1925 he made a lecture tour in the US. See Dorrien, *Barthian Revolt*, 108, 111; Adams, "Theological Ethics of the Young Rawls," 29.
[40] Reinhold Niebuhr, *The Nature and Destiny of Man. Volume I: Human Nature* (New York: Charles Scribner's Sons, 1964 [1941]), 127-8, 132.

theologian Anders Nygren's *Agape and Eros*, published in two volumes in 1930 and 1936 and translated into English between 1932 and 1939.[41] Rawls read *Agape and Eros* and, as we will see, used it in his arguments.[42] Contrary to Adams and Reidy's suggestions, Nygren was not a neo-orthodox but a thorough-going biblical historicist.[43] He wrote a dissertation on the biblical historicist Ernst Troeltsch [1865-1923], and in his famous 1922 essay, "The Essence of Christianity," repeated the core historicist theme that, despite the various forms Christianity took in different social contexts, the historical figure of Jesus Christ was the uniting link.[44] To recover the lost essential meaning of Christiantiy, Nygren looked at how it defined itself against its rivals, in particular the Greek idea of love.[45] Thus the idea of "essence" in Nygren's work differed from the mentioned ones and was akin to "its original, untainted expression." This conception of the essence of Christianity would be influential on Rawls.

Thus Biblical essentialism persisted in American theological thought by merging the neo-orthodox themes into its framework. Yet, it persisted without some of the characteristics which defined the German historicism. Theologians continued to assume that the Christian experience is shared, they did not think that historical development changed the essence of Christianity.[46] This absence of historicism is particularly evident in the writings of Theodore M. Greene and George F. Thomas, two Princeton University professors who were to shape Rawls's early commitments. Their arguments shed some light on Rawls's early religious commitments and his later uneasy relationship with historical approaches to political philosophy. Greene, professor at the philosophy department and one of the supervisors of Rawls's thesis, understood God as a person who revealed himself in more than one way: through the historical figure of Jesus, in the natural world, but also "in the distinctive religious experiences of mystics, saints and prophets, and, more particularly, in the individual and corporate experiences recorded in the Bible."[47] As was typical of the neo-orthodox, Greene allowed that God can be known through the Bible and Christ, but his overall conception of the Bible drew on the historicist and essentialist themes: it was not the standard of truth, but a record of Christian experiences. Like biblical essentialists, Greene placed the weight of his argument on the more direct religious experiences of the Christian community: the trust in the truthfulness of these experiences was warranted, he thought, because they were shared by all Christians.[48]

---

[41] Anders Nygren, *Agape and Eros*, Philip S. Watson, trans. (Philadelphia: The Westminster Press, 1953 [1932-1939]).
[42] Rawls, *Meaning of Sin and Faith*, 135, 174, 254.
[43] Adams, "Theological Ethics of the Young Rawls," 31; Reidy, "Rawls's Religion," 315.
[44] Thor Hall, *Anders Nygren* (Waco, TX: Word Books Publishers, 1978), 36; Anders Nygren, *The Essence of Christianity: Two Essays*, P. S. Watson, trans. (London: The Epworth Press, 1960), 12, 57. Nygren's original argument was that *agape* had undergone a series of changes since its inception, especially when contaminated by the Greek concept of love, *eros*. Due to these contaminations, it was no longer evident what the "essential meaning" of *agape* was, and the task of theology was to clarify it.
[45] Nygren, *Agape and* Eros, 29-30.
[46] Some of the early American biblical essentialism was, on the contrary, explicitly historical. For example, William Adams Brown argued that, "If Christianity is to make good its claim to be absolute religion, it must be able to show that it is the goal of religious progress." Brown, *The Essence of Christianity*, 290-291. However, he rejected the Hegelian account of history as too simple to explain the facts. See ibid, 304.
[47] Theodore M. Greene, "Christianity and Its Secular Alternatives," in Paul J. Tillich et al., eds., *The Christian Answer* (London: Nisbet & Co., 1946), 101.
[48] Greene, "Christiantiy and Its Secular Alternatives," 104.

Professor of religious thought George F. Thomas, whose class on the Christian ideas to the Reformation Rawls attended, was, much like Greene, a thinker influenced by both neo-orthodox and biblical essentialist movements. God, Thomas wrote, discloses himself to man, and does so as a person to a person.[49] Since revelation is not a straightforward passing down of truths but a personal encounter, the Bible, Thomas admitted, could contain errors.[50] He argued that to understand the Christian faith we need not so much an intellectual interpretation, but attention to the Christian experience.[51] Although Thomas did not think that this experience exhibited a commonality in all of its expressions, he nonetheless relied on Christian experience – and not the Bible – as the justification for his Christian teaching. Thus, while already displaced as a movement, biblical essentialism – now modified by some neo-orthodox commitments – still informed American philosophers and theologians in 1939. But, in the form it reached the young John Rawls, it had lost its connection with historical inquiry: Harnack's insistence that one should look at history to uncover the essence of Christianity no longer informed Rawls's teachers.

Meaning of Sin and Faith

Growing up in this mixed landscape of American Protestantism, guided by his biblical essentialist teachers, Rawls would write *Meaning of Sin and Faith*. While not consciously taking positions in the debate about the basis of Christian universalism, the thesis was remarkably biblical essentialist in character. Its main questions sprung from typically biblical essentialist themes, in particular belief in the commonality of Christian experience, reliance on this experience as the ultimate source of appeal, and the belief that the theologian and philosopher had the task of analyzing this shared Christian experience. Guided by these commitments, Rawls undertook to analyze Christian experience by outlining a conceptual framework that explained it.[52] Along the way, he argued that the most prevalent attempt at such an analysis – "naturalism," as he called it – was gravely mistaken. His main innovation – the extension of the shared experience from Christians to the entire humanity – also sprung from the biblical essentialist framework. Thus, while recent commentators are certainly right to highlight neo-orthodox themes in the thesis – the emphasis on the knowledge of God as a personal disclosure, the impossibility of knowing God through reason alone – they are mistaken to call it a neo-orthodox work. Neo-orthodox themes inform some aspects of Rawls's description of the common Christian experience; they are needed to explain Rawls's eventual turn to secularism, but they do not motivate his main questions in the thesis.[53] To explain these, we need to appeal to biblical essentialism, especially its de-historicized 1940s form.

Rawls's main goal in the thesis was typical of biblical essentialism: to elaborate a conceptual framework that analyzed our experience correctly. "Every theology and every philosophy proceeds to investigate experience," he wrote, and it does so "with certain

[49] George F. Thomas, "Central Christian Affirmations," in Paul J. Tillich et al., *The Christian Answer* (London: Nisbet & Co., 1946), 130, 136.
[50] Thomas, "Central Christian Affirmations," 132, 138.
[51] Thomas, "Central Christian Affirmations," 180.
[52] Rawls, *Meaning of Sin and Faith*, 110.
[53] See footnotes 1 and 2 of this chapter.

fundamental presuppositions."[54] Following the aims of philosophy and theology, Rawls proposed to "outline and investigate our fundamental presuppositions."[55] Sketching four core presuppositions – the existence of God, personality, community, and the distinctness of these three from what he called the natural world – he claimed that these and the attendant notions explained the shared Christian experience best.[56] As many theologians and philosophers of his time, in his analysis of Christian experience Rawls wove neo-orthodox themes into the overall biblical essentialist framework. This is particularly true of his conception of God – the core of this conceptual framework.

God, Rawls thought, was a person. Although he claimed that being a person is not reducible to mental states, his examples of personhood turned precisely on such mental states: the constitutive characteristic of personhood was the ability to recognize others as persons.[57] This, in turn, expressed itself in such relations as being able to love others, give to them, share with them, despise them, and, on the whole, behave toward them differently than toward objects.[58] From the examples Rawls gave, we can gather two main characteristics of personal relations and therefore gain insight into God's person. First, personal relations were governed by thought-out decisions that stemmed from a certain understanding of proper behavior toward others. They were not governed by what he called natural impulses; thus, Rawls contrasted personal relations with "natural" relations, or those driven exclusively by appetitions and desires.[59] This latter, natural, world was, according to Rawls, merely an "expanse of space filled by bodies, all that we see, feel, touch and so forth." [60] The world governed by personal relations, on the other hand, implied an ethical community. Second, personal relations implied acknowledging another's individuality: her goals and desires, and, more broadly, her person. Following Brunner and the neo-orthodox tradition, Rawls thought that all knowledge of another person comes through that person's decision to disclose herself: "all knowledge of other persons is knowledge given to us by them."[61] Thus, necessarily, personal relations are always "active on both sides," they always proceed on the basis of "mutual self-revelation."[62] For that reason, Rawls thought that personal relations are "unique" in that the partners of this conversation or mutual encounter are not "readily exchangeable": an encounter with a different person would be a different encounter.[63] As he later explained using an example of a natural relation – prostitution – in natural relations "other people can only enter into [one's] consciousness as means to the achievement of the desired end"; it makes no difference who or what the person is: each is "as good as another."[64]

---

[54] Rawls, *Meaning of Sin and Faith*, 110.
[55] Rawls, *Meaning of Sin and Faith*, 110.
[56] Rawls, *Meaning of Sin and Faith*, 111-13.
[57] Rawls, *Meaning of Sin and Faith*, 113, 148.
[58] Rawls, *Meaning of Sin and Faith*, 180.
[59] Rawls, *Meaning of Sin and Faith*, 112, 117. Rawls understood "appetition" as a striving that does not create any personal relation between the subject and the object of appetition. Ibid., 117.
[60] Rawls, *Meaning of Sin and Faith*, 112.
[61] Rawls, *Meaning of Sin and Faith*, 224.
[62] Rawls, *Meaning of Sin and Faith*, 115, 117-18.
[63] Rawls, *Meaning of Sin and Faith* 117.
[64] Rawls, *Meaning of Sin and Faith*, 187.

God, according to Rawls, was a person in both of these respects. First of all, he was a paragon of personal or communal relations, as He was "in Himself [a] community, being the Triune God."[65] In fact, Rawls thought, God created human beings in His own image, with the establishment of community in mind; thus, the capacity for community was the best in humans.[66] Second, Rawls wrote, God was a person in that he, like other persons, revealed his own person in personal encounters of mutual self-revelation. Thus, like the neo-orthodox, he thought that almost all knowledge of God comes from his decision to reveal himself: "man must wait for God to speak to him. He must wait for His word."[67] God's self-revelation, as Adams and other commentators point out, is a typical neo-orthodox theme. However, Rawls did not, unlike the neo-orthodox, think that God revealed himself only once, through Christ; rather, he thought that God also disclosed himself in direct, personal, and rather frequent encounters with human beings.

Archetypical of such personal encounters with God's Word was the experience of conversion, which Rawls described as "that intense experience of flatness and lying in exposure before the Word of God."[68] As Adams points out, Rawls's conception of the Word of God – God's disclosure – is very broad: it includes God's incarnation in Jesus, the working of the Word through the chosen apostles who spread Jesus's story, and – crucially – in the personal experiences of conversion as described in the Acts of the Apostles and by Rawls in his thesis.[69] Of these, Rawls emphasized the latter: conversion, he thought, "constitutes the synthesis of Christian experience," and all doctrines of election "which do not spring straight from it are purely academic."[70] Adams remarks that this "general lack of emphasis on christology" and the emphasis on the converting activity is "perhaps the least neo-orthodox feature of Rawls's treatment of revelation."[71] That is certainly right – the personal experience of God in conversion is a biblical essentialist and generally liberal Protestant theme.

Biblical essentialism also helps explain Rawls's reliance on these personal encounters to justify his analysis of the shared Christian experience. Although Rawls claimed that conversion was "a formless and indeterminate kind of experience," he believed that, far from defying reason, it provided us with knowledge of God.[72] As we have seen, most of Rawls's conceptual framework relied on the concept of a person and the corresponding distinctively human capacity to be a member of a community. Experiences of conversion helped Rawls justify such statements: they clarified, among other things, that "[God] wants a community bound together in faith and rejoicing in thanksgiving" and that he reveals himself to achieve it.[73] As Rawls thought that "the nature of God … is …represented to us unmistakably in the experience of His

[65] Rawls, *Meaning of Sin and Faith*, 113.
[66] Rawls, *Meaning of Sin and Faith*, 112, 203.
[67] Rawls, *Meaning of Sin and Faith*, 224. Rawls also allowed that by reason we can learn that God is intelligent, powerful and eternal, but he denied that reason can get us any further than that. Ibid.
[68] Rawls, *Meaning of Sin and Faith*, 223.
[69] Adams, "Theological Ethics of the Young Rawls," 96; Rawls, *Meaning of Sin and Faith*, 124-5, 233-4.
[70] Rawls, *Meaning of Sin and Faith*, 233.
[71] Adams, "Theological Ethics of the Young Rawls," 96.
[72] Rawls, *Meaning of Sin and Faith*, 234.
[73] Rawls, *Meaning of Sin and Faith*, 241.

Word," such experiences helped him arrive at his understandings of "person" and the content of *Imago Dei*.[74]

This biblical essentialist theme needs to be emphasized, as it defines Rawls's work: he thought that his conceptual framework needed no further support than showing that Christians actually had these experiences. As he wrote, his main commitments "have empirical meaning and are derived from experience," such as the discussed example of conversion.[75] This theme explains why, despite remarking in the bibliography that the Bible is "always the last word in matters of religion," Rawls did not think that the truth of his conclusions depended on their correspondence to statements in the Bible.[76] Like his teacher Theodore Greene, Rawls understood the Bible as a record and analysis of Christian experiences. As such, the Bible was a source of examples, some of which – such as Peter's speechlessness or Paul's being struck dumb – Rawls used as examples of conversion similar to his own.[77] He did think that the Bible, narrating the experiences of conversion, fully revealed the nature of God: "The Bible has told us all we need to know about Him."[78] He also certainly thought that his conclusions about God coincided with those in the Bible; as he wrote in the introduction to the thesis, his analysis of the Christian experience was "a rehash of what everybody knows."[79] However, he did not think that the Bible was the standard of truth, and, in fact, appraised it by this very same standard: "the Bible is right," he wrote, "when it insists that we will be resurrected in some sort of body, whatever sort it may be."[80]

To make the reliance on Christian experience good, Rawls had to show that it was actually shared. For this reason, he insisted that all Christians, although they would experience conversion in different ways – some suddenly, others in a protracted way – would agree on its content.[81] "If any of us analyze our experience, and if that experience is genuinely Christian," he wrote, "then we should all agree. Although we may have never experienced a sudden conversion like Paul's, we can nevertheless understand Paul and agree with him."[82] As he thought that all Christians had experienced conversion and that this experience provided knowledge of God, Rawls expected that they would all recognize the conceptual framework elaborated in *Meaning of Sin and Faith* as their own.

Rawls's reliance on this dehistoricized biblical essentialist theme involved him in a dilemma that would occupy him up until the publication of *A Theory of Justice*: how to explain the possibility and nature of disagreement in light of the allegedly shared experience. Clearly, not all Christians, let alone everyone, would have agreed with his analysis of the shared experience. To answer this dilemma, Rawls put forth a distinction between experiences and theories that analyze this experience. Thus, while all persons could be expected to have at least some of the same experiences, not all persons would agree on theories that analyze these

[74] Rawls, *Meaning of Sin and Faith*, 243.
[75] Rawls, *Meaning of Sin and Faith*, 113.
[76] Rawls, *Meaning of Sin and Faith*, 254.
[77] Rawls, *Meaning of Sin and Faith*, 236.
[78] Rawls, *Meaning of Sin and Faith*, 111.
[79] Rawls, *Meaning of Sin and Faith*, 110.
[80] Rawls, *Meaning of Sin and Faith*, 153.
[81] Rawls, *Meaning of Sin and Faith*, 234-5.
[82] Rawls, *Meaning of Sin and Faith*, 234-5.

experiences. It is interesting that Rawls's strategy in answering this dilemma in 1942 was not only to show the rival analyses – rival theories – wrong but also to appeal to a historical analysis to show just where and when they went wrong. Thus, while *Meaning of Sin and Faith* as a whole did not proceed historically, it contained uses of historical analysis common to biblical essentialists.

To retain his key commitment that all Christians have at least some of the same experiences, Rawls first argued that Augustine and Aquinas shared the Christian experiences but failed to analyze them correctly.[83] According to Rawls, Augustine and Aquinas overlooked the crucial distinction between the personal and the natural: "all naturalistic thinkers have completely missed the spiritual and personal element which forms the deep inner core of the universe."[84] To prove their analyses wrong, Rawls attempted to show that natural appetitions cannot lead to personal relations, either to egotism (the worst type of personal relation) or to community (the proper personal relation). Since, he assumed, we all had experiences of egotism and communal relations, this argument showed that Augustine's and Aquinas's concepts did not analyze our experiences correctly. For this failure to appreciate the distinction and to describe all human relations in terms of desires, he named Augustine, Aquinas and thinkers failing to make the distinction between natural and persons relations "naturalists."[85]

Rawls also pointed out the historical origin of Augustine's and Aquinas's naturalist mistake. To do so, he appealed to the historical themes of biblical essentialism. Using the method of Nygren, he traced Augustine's and Aquinas's mistake to the Greeks – Plato and Aristotle – who analyzed all human relations in terms of desire and appetition. All these thinkers, Rawls argued, thought that the task of ethics is to turn human desire toward a proper object; Augustine and Aquinas only changed that proper object to God.[86] In doing so, he thought, they turned God into "merely a bigger and better object of … enjoyment."[87] To the contrary, Rawls believed that ethics was not about desire but about personal relations: "proper ethics is not the relating of a person to some objective "good" for which he should strive, but is the relating of a person to person and finally to God."[88] On the whole, then, Rawls thought that his belief in the commonality of Christian experience did not require actual agreement, as long as one could show why the rival analyses fail and in what ways they depart from Christianity's original and proper expressions.

In other respects, Rawls departed from liberal Protestantism's tenets. Most importantly, he expanded the shared experience from Christians to non-Christians and even non-believers. To Rawls, this step was quite self-evident: he thought that non-believers would have experiences typical not of faith, but of sin. In particular, he wrote, sin would engender the feeling of aloneness, or "spiritual cut-offness" and "desolating closedness."[89] Nietzsche, in

[83] This view can be seen in his choice of words: Rawls wrote that the rival view of sin "does not accord with the facts," that "examination of our actions bears him out," and that the concept of "will" is "a false representation of personal experience." See Rawls, *Meaning of Sin and Faith*, 191, 161, 220.
[84] Rawls, *Meaning of Sin and Faith*, 120.
[85] See, for example, Rawls, *Meaning of Sin and Faith*, 161-2.
[86] Rawls, *Meaning of Sin and Faith*, 120-1.
[87] Rawls, *Meaning of Sin and Faith*, 162.
[88] Rawls, *Meaning of Sin and Faith*, 161, 114.
[89] Rawls, *Meaning of Sin and Faith*, 122-3, 206.

Rawls's mind, was a good example. Quoting a passage in which he had proclaimed that "this world is the *Will to Power* and nothing else," Rawls concluded that Nietzsche's world "is one of aloneness" and, as aloneness is one of the experiences of sin, that his experiences are best described as experiences of the sinful.[90]

The example of Nietzsche shows how peculiar Rawls's reach for non-Christians was: he allowed that a person who had not thought of God and did not think of his own actions as sinful would still have experiences typical of sin. He did not compare Nietzsche's experiences of sin with those of a repenting Christian who thought of himself as sinful, in effect detaching these experiences from the beliefs which typically give rise to them. As Rawls did not think that such a move needed an explanation, we have no explicit statements regarding it. However, various dispersed remarks point us to *Imago Dei* and experiences emerging from a structural relation with God. In virtue of this image, Rawls thought, humans always have the capacity for community.[91] This capacity can be rejected, but the *Imago Dei* can never be abrogated: as he wrote, "all men have God as Father, but not all men are His sons."[92] Because of this image, he thought, humans are always in a certain relation to God, their actions can always be described as sinful or faithful, and – Rawls seems to have concluded – a repudiated relation to God creates experiences of sin: "Aloneness is aloneness because the *Imago Dei* remains."[93]

In sum, to understand the driving questions of *Meaning of Sin and Faith* correctly, we need to note both neo-orthodox and biblical historicist and biblical essentialist themes, but we must emphasize that the main goals of the thesis stem from commitments typical of the latter. Rawls did depart from this tradition by extending the commonality of experience from Christians to everybody. However, biblical essentialist themes explain why Rawls saw his task as the analysis of experience, why he relied on the more direct and personal experiences for support of the results of his analysis, as well as why he provided no further argument for his commitments than their reliance on universally acknowledged experiences. Moreover, Rawls's commitment to biblical essentialist themes also helps to explain his dilemmas at the end of the Second World War and the main features of his resulting secular thought. As we will see, war events forced him to question the possibility of direct and personal experiences of God, on which the conceptual framework of *Meaning of Sin and Faith* relied.

The Second World War and the Turn to Secularism

In February 1943, having finished his thesis and undergraduate education, Rawls enrolled in the Army. Sent to the Pacific theater for two years, he served in New Guinea, the Philippines, and, toward the end of the war, in Japan. Overall, he judged the army a "dismal institution," left it in January 1946 and re-entered Princeton University in the autumn of the same year, this time as a graduate student in philosophy.[94] Within a year, he was a changed person: not only did his

[90] Rawls, *Meaning of Sin and Faith*, 212, 213 (emphasis original).
[91] Rawls, *Meaning of Sin and Faith*, 121.
[92] Rawls, *Meaning of Sin and Faith*, 244.
[93] Rawls, *Meaning of Sin and Faith*, 208.
[94] Thomas Pogge, *John Rawls: His Life and Theory of Justice*, Michelle Kosch, trans. (Oxford: Oxford University Press, 2007), 12.

arguments not rely on God or other Christian concepts that structured *Meaning of Sin and Faith*, but he did not even mention these concepts until 1951 when, already an instructor at Princeton, he reviewed Princeton theologian Paul Ramsey's *Basic Christian Ethics*.[95] As Rawls did not reflect on his Christian past in his graduate school writings, we do not have any contemporaneous documents that reveal his dilemmas. He did, however, write reminiscences on his religion and the years in the army in the early 1990s; they are published together with *Meaning of Sin and Faith*, and provide us with three events during the Second World War that aid in elucidating the transformation in his thinking.

All mentioned events question the feasibility of his biblical essentialist conception of the personal and direct experience of God, which, as we have seen, depended on God' action – personal relations, he wrote, were always "active on both sides" – and, as personal relations were not intercheangeable, on God's action toward particular persons.[96] The first of these events is the speech of a Lutheran pastor at Kilei Ridge in December 1944. Encouraging the soldiers before battle, the priest proclaimed that God directed the American bullets at the Japanese and protected the Americans from the bullets of their enemies. Rawls judged these claims as "simply falsehoods," yet these falsehoods made him question his own understanding of God.[97] Rawls had combined neo-orthodoxy and biblical essentialism by claiming that God reveals himself personally, and not only through Christ, but also in the more direct experiences of conversion known to every Christian. Numerous deaths in the war questioned this picture of God as someone who frequently intervenes into human affairs – even if only by self-disclosure – and led Rawls to the conclusion that God was disengaged from the human world.

The death of Rawls's tent-mate and friend Deacon must have made this conclusion very apparent. In May 1945, on the Villa Verde trail on Luzon, Deacon died entirely due to what Rawls saw as the chance of circumstances. When the First Sergeant asked for two volunteers, one to reconnoiter the Japanese position and the other to give blood to a wounded soldier, Deacon and Rawls agreed that the tasks would depend on their type of blood. Rawls's was appropriate while Deacon's was not; so Deacon went to reconnoiter and, hit with a mortar shell, died.[98] Rawls could not give this death higher purpose, and God appeared more and more withdrawn from details of human life. The third event, the news about the Holocaust from the first American troops to reach the German concentration camps, strengthened this conclusion. While, on his own account, Rawls had gone along with Lincoln's attempt to give the Civil War purpose and paint God as acting justly, the Holocaust, Rawls wrote, "can't be interpreted in that way."[99] Realizing that God would not intervene to save millions of Jews, he concluded that he could not expect God's response to prayer or any intervention into human affairs.[100] As his conception of God's personal self-disclosure in the experiences of conversion implied an active God that disclosed himself to particular persons, Rawls must have rejected it for the same reasons.

---

[95] Paul Ramsey, *Basic Christian Ethics* (New York: Charles Scribner's Sons, 1950).
[96] Rawls, *Meaning of Sin and Faith*, 115, 117-18.
[97] John Rawls, "On My Religion," in Rawls, *Meaning of Sin and Faith*, 262.
[98] Rawls, *Meaning of Sin and Faith*, 262.
[99] Rawls, *Meaning of Sin and Faith*, 262.
[100] Rawls, *Meaning of Sin and Faith*, 262.

Rawls's war experiences by no means necessitated the abandonment of religion altogether: he could have simply modified his understanding of God as well as the accompanying notions of sin, faith, and revelation. In fact, like many biblical historicists and essentialists during the First World War, he could have turned to neo-orthodoxy's conception of revelation as God's disclosure through Jesus, rejecting his account of personal contact with God's Word in conversion. Yet, he did not. In his later writings he dropped all these concepts, which suggests that he lost his faith. In his reminiscences, however, Rawls wrote that during "the following months and years" he rejected many Christian doctrines but that his "fideism remained firm against all worries about the existence of God."[101] This memory implies that, contrary to my initial suggestion, Rawls's faith in God remained untouched. Likely, as did many intellectuals and academics of the period, he thought of himself as in other respects a Christian, but not sufficiently such to use the typical Christian concepts in his arguments. We do not have any other contemporary writings to shed further light on the nature of the dilemmas that drove Rawls to what was in all relevant respects an atheist position. Perhaps he could not conceive of a God who was simultaneously just and who allowed the Holocaust to happen. Lacking this concrete picture of God, Rawls did not have a criterion to distinguish between the just and unjust. This dilemma would explain part of Rawls's criticism of Paul Ramsey's *Basic Christian Ethics*. Alternatively, perhaps different cultures of the Pacific theater made Rawls realize that Christianity could not explain experiences of all human beings, and, as a result, made him seek another basis for his universalism.

While the dilemmas of the Second World War led Rawls to reject his picture of God as well as reliance on personal experience of God's self-revelation, other themes common to de-historicized biblical essentialism continued to influence his thinking. In particular, Rawls continued to hold his earlier – although now secular – beliefs about the task of theology and philosophy. Like that of American biblical essentialists, it was a conception that had no particular place for history and historical understanding. Rawls thought that Christians and non-Christians had shared experiences and that a theologian or philosopher analyzes this shared experience. His earlier belief that to justify a theological theory one only needed to show that it analyzed shared experience correctly also continued to look convincing. And finally, his belief that no further argument than this reliance on agreement needed to be made would keep informing his later views of philosophy.

Conclusion

As a theological work, Rawls's *Meaning of Sin and Faith* is a reflection of the battles that took place in American theology between biblical historicism and essentialism and neo-orthodoxy. If neo-Hegelian biblical historicism failed to have influence in 20th century America, by the time Rawls wrote his thesis the related biblical essentialism had lost its connection to historical research. This character of American theology reached Rawls through his teachers and thesis supervisors Theodore Greene and George Thomas. As a result of these historical developments in the 20th century, Rawls's thesis has the character of an empiricist philosophy, in which theology has a task of constructing a theory that explains shared experiences. This

[101] Rawls, *Meaning of Sin and Faith*, 262.

conception of theology is importantly anti-foundational: it does not require to justify a theory by grounding it in certain knowledge. And it is also importantly ahistorical, holding implicitly that experiences, even if they stem from different conceptual backgrounds, are shared.

This de-historicized biblical essentialist conception of philosophy helps explain Rawls's shift to a secular positivist philosophy. As Rawls casted away his religious beliefs and rejoined Princeton as a graduate student in 1946, he worked under the supervision of the empiricist philosopher Walter T. Stace. Drawing on the remnants of his biblical essentialist conception of philosophy, Rawls turned to the dominant philosophical tradition of the time: logical positivism.

3

The Early Positivist Years

Introduction

It is often thought that ethical inquiry was revived in the 1960s and 1970s, and that this revival took place against an ethical background laid waste by logical positivism.[1] Rawls is one of the principal actors in this popular narrative; he is shown to have brought back a more classic and also more fruitful approach to moral and political philosophy. While this popular narrative overstates the barrenness of the 1950s and 1960s and exaggerates Rawls's role in said revival, there is nonetheless much truth in the contours of its story.[2] In the 1940s and 1950s, Anglophone moral philosophy was dominated by two ethical currents of logical positivism, neither of which viewed ethical reasons as worthy of consideration. Naturalism thought them reducible to statements about empirical facts, while emotivism treated them as expressions of emotion not susceptible to rational consideration. Rawls's articles in the 1950s and the 1960s offered a new conception of objectivity in ethics and made possible normative discussion that rested on reasons.

However, while this popular narrative claims that Rawls rejuvenated Anglophone political philosophy by borrowing from the social contract tradition, archival research shows that he in fact drew on logical positivism, modeling ethics on the tradition's conception of science. While initially surprising, this intellectual connection is intelligible for three important reasons. First, logical positivism was widespread in Anglophone philosophy. When Rawls rejoined the Princeton philosophy department as a graduate student in the Spring of 1946, logical positivism's conception of philosophy as a scientific enterprise and the attendant themes of the empiricist criterion of meaning, non-foundational justification, formalism, and disregard of the intentional meaning were ubiquitous, all the more because pragmatism, which also saw itself as a scientific and empiricist philosophy, eventually adopted many of these commitments.[3] Second, logical

[1] This popular narrative originates mainly from Peter Laslett's provocative assessments of the state of political philosophy in the 1950s. See his introductions to *Philosophy, Politics and Society*, ii, ix, and to Peter Laslett and James Fishkin, eds. *Philosophy, Politics and Society*, Fifth series (New Haven: Yale University Press, 1979), 2.

[2] For criticisms of this popular narrative, see Robert Adcock and Mark Bevir, 'The Remaking of Political Theory' in Robert Adcock, Mark Bevir and Shannon Stimson, eds. *Modern Political Science: Anglo-American Exchanges since 1880* (Princeton, 2007), 209-233; Petri Koikkalainen, 'Peter Laslett and the Contested Concept of Political Philosophy', *History of Political Thought* 30 (2009): 336-359.

[3] For logical positivism's relationship with American pragmatism, see Alan W. Richardson, "Logical Empiricism, American Pragmatism, and the Fate of Scientific Philosophy in North America" in Gary L. Hardcastle and Alan W. Richardson, eds. *Logical Empiricism in North America* (Minneapolis, MN: University of Minnesota Press, 2003), 1-24, and Cornelius Delaney, "Realism, Naturalism, and Pragmatism," in Thomas Baldwin, ed. *The Cambridge History of Philosophy 1870-1945* (Cambridge: CUP, 2003), 449-460. Both Richardson and Delaney emphasize scientific aspirations of the two traditions, but it also needs to be added that in the mid- to late-1930s, when logical positivism had become non-foundational, reliance on experience had become very similar in the pragmatist and the logical positivist traditions, and was perceived so by the representatives of these traditions. See, for example, Herbert Feigl's "Method Without Metaphysical Presuppositions" in *Philosophical Studies* 5 (1954): 17-29, in which

positivism was a variegated tradition both in philosophy more broadly and in ethics. Equating logical positivism with A.J. Ayer's emotivism, we fail to note the tradition's attempt to construct an empirical yet normative theory in ethics. Skirting logical positivism's dichotomy between analytic and synthetic truths but nonetheless relying on its conception of scientific inquiry, this current is better named positivist, not logical positivist. Hence the name for Rawls's conception of philosophy: it was positivist despite its origins in logical positivism. Lastly, Rawls's earlier biblical essentialist conception of philosophy had many connections with logical positivism, even if this latter was generally deeply opposed to religious views. Both assumed underlying agreement in everyone's experiences, both tried to elaborate theories to explain these shared experiences, both were very ahistorical and both did not think that one could provide a more foundational justification for an ethical theory. Thus the shift from biblical essentialism to logical positivism was not far-fetched; indeed, it helped Rawls preserve and elaborate on commitments crucial to his earliest conception of philosophy.

In this chapter, I want to narrate the formation of Rawls's conception of philosophy between 1946 and 1951. To do that, I will first outline the core themes of logical positivism: its non-foundational reliance on experience, deductive structure of scientific theory, empiricist criterion of meaning and the resulting disregard of the intentional meaning of scientific and ethical judgments. In the second part, I will outline the three logical positivist currents in ethics: naturalism, emotivism, and scientism – the mentioned new, most promising, current. In the third part, I show how, inspired by this third logical positivist current, Rawls developed his own "physicalist" theory of ethics. Its core idea was to model ethical inquiry after empirical inquiry and claim that it differed from scientific inquiry only in its subject matter. Following the analogy, if observational statements were the subject matter of a scientific theory, ethical judgments played this role for an ethical theory. In other respects the two types of inquiry were the same. Both aimed at building a deductive structure consisting of axioms and the basic observational statements deduced from these axioms. Theories in both types of inquiry were justified by the ability of their axioms to deduce observational statements (OS_D) that matched the actual observational statements (OS_A) made by the scientific community. And neither scientific nor ethical theories could be justified in a more foundational way.

In the initial years, Rawls's positivist conception of philosophy proved fruitful, and he spent much of 1946 and 1947 developing it. I outline these developments and modifications in the fourth part of the chapter. Focusing on Rawls's central task of detailing the account of objectivity of ethical judgments, I explain his claim that ethical judgments are considered objective insofar as "normal observers" agree on them. This conception of objectivity would set Rawls's goal for this period: he would attempt to show that all "normal observers" do in fact make at least some of the same judgments. He would develop the notions of the "reasonable man" and "rational judgment" as part of the attempt to show this agreement.

While Rawls drew on logical positivism in constructing the outlines of his conception of philosophy, he was also influenced by other intellectual traditions. I outline the first motifs of linguistic philosophy in Rawls's thought. As Rawls drew only on its notion of "absurdity" but left aside the wider conceptual nexus which gave it significance, linguistic philosophy had little

---

he notes that pragmatism's most valuable contribution to epistemology was its claim that one can only vindicate (as opposed to validate by proof) theories by showing their usefulness to human purposes (Ibid, 26).

systematic influence on him between 1947 and 1950. Nonetheless, even in this period linguistic philosophy led Rawls's interest away from uses of ethical words to their intentional meanings: he stopped being interested in why one made ethical judgments and started focusing on their content. As a result, his ethical theory became noticeably less formalistic but, unlike *A Theory of Justice*, it was still concerned not with the picture of a just society but with the most general conceptual connections which form the background of any reasonable person's view of the world.

As logical positivism set Rawls's goals, so it created his dilemmas, some of them immediate and some that would surface only in later years. These dilemmas would shape his thinking even when he would reject some of the commitments that created them. I outline these dilemmas in the fifth part of the chapter, focusing on Rawls's most immediate inability to explain the agreement of reasonable persons without retreating to moral realism. This dilemma would reveal the shortcomings of the analogy between ethics and science and eventually push Rawls away from it.

Logical Positivism

Logical positivism's overall philosophical approach consisted in combining the earlier positivist tradition with the recent discoveries in logic. It developed in Vienna in the 1920s from many influences but particularly from the positivist tradition of Ernst Mach and the logical school of Gottlob Frege, Bertrand Russell, and the early Ludwig Wittgenstein. This combination of positivism and advances in logic was reflected in the tradition's key themes: its empiricist reliance on experience, deductive conception of scientific theory, and the criterion of meaningfulness which claimed that all but empirical (synthetic) and logical (analytic) statements are pseudo-statements or simply "nonsense." I want to explain these themes and outline the evolution of logical positivism to the form that would be most influential on Rawls.

In our popular narratives, logical positivism is associated with foundationalist interpretations of experience. This characterization well describes the early works of the tradition, such Rudolf Carnap's [1891-1970] *The Logical Construction of the World* or Moritz Schlick's [1882-1936] "The Turning Point in Philosophy."[4] These early logical positivists thought that basic or elementary experiences are not permeated by the conceptual frameworks of the people who have them. Not depending on any conceptual framework, such experiences could be used to justify these frameworks. Thus, taking these experiences as a "given," Carnap aimed to construct a conceptual structure that rested on these experiences, while Schlick proposed a similarly foundational structure, arguing that, to justify a proposition, we must in the end resort to non-linguistic "pointings, in exhibiting what is meant."[5]

By the mid-1930s, however, this form of logical positivism was replaced by a non-foundationalist positivism which interpreted all experience in linguistic terms: all experience was

[4] Rudolf Carnap, *The Logical Structure of the World: Pseudoproblems in Philosophy*, trans. Rolf A. George (Berkeley, CA: University of California Press, 1969); Moritz Schlick, "The Turning Point in Philosophy" trans. David Rynin In *Logical Positivism*, ed. A.J. Ayer (Westport, CT: Greenwood Press, 1959), 53-59.
[5] Carnap, *Logical Structure of the World*, 7, 19, 102; Schlick, "The Turning Point in Philosophy", 57.

now thought to depend on conceptual schemes or "descriptions" and "classifications."[6] Carnap too left the foundationalist camp, concluding that his earlier notion of experience was unworkable.[7] Logical positivists now thought that, lacking a rock-bottom foundation in non-linguistic experiences, all propositions were justifiable only by other propositions of the same kind. This changed the nature of justification drastically: in principle, it went on forever and, when it stopped, it stopped in a tentative manner and for pragmatic reasons. As Karl Popper [1902-1994], one of the key figures of the movement and a direct influence on Rawls, made it clear, justification stopped when the scientific community agreed on a sufficient number of observations to declare any one scientific theory correct. In one of his most eloquent passages, Popper compared the construction of scientific theories to building in swamps:

> The empirical basis of objective science has thus nothing 'absolute' about it. Science does not rest upon rock-bottom. The bold structure of its theories rises, as it were, above a swamp. It is like a building erected on piles. The piles are driven down from above into the swamp, but not down to any natural or 'given' base; and when we cease our attempts to drive our piles into a deeper layer, it is not because we have reached firm ground. We simply stop when we are satisfied that they are firm enough to carry the structure, at least for the time being.[8]

While logical positivism changed significantly due to this shift to non-foundationalism, it nonetheless preserved some of its earlier features – features that would influence Rawls. In particular, logical positivists maintained the belief that some experiences are basic, and not merely more basic than other experiences. The notion of "observational statements" (also called "basic" or "protocol" statements") for which logical positivism is known is directly tied to these basic experiences.[9] Observational statements were typically thought of as records of these simple experiences: they were "self-consistent singular statements" of fact that reported "observable events" occurring at a given time and a given place.[10] The notion of basic experience led logical positivists to the belief that, as long as typical scientific observers are appropriately placed with regard to the object of observation, they would agree in their reports. Popper took agreement among scientific observers as a matter of course: his notion of "observation," albeit not elaborated, only required that the observers be "suitably placed in space and time."[11] Indeed, Popper was so convinced of this agreement that he concluded that, should it prove impossible, it

---

[6] Alfred Jules Ayer, *Language, Truth, and Logic* (New York: Dover, 1946), 91. For a brief yet excellent account of the development of logical positivism from foundationalism to non-foundationalism, see Carl G. Hempel, "On the Logical Positivits' Theory of Truth," *Analysis* 2 (1935): 49-59. Also valuable is Ernest Nagel's "Impressions and Appraisals of Analytic Philosophy in Europe. I" *The Journal of Philosophy* 33 (1936): 5-24 and "Impressions and Appraisals of Analytic Philosophy in Europe. II" *The Journal of Philosophy* 33 (1936): 29-53, and Thomas Uebel, "Anti-Foundationalism and the Vienna Circle's Revolution in Philosophy", *British Journal for the Philosophy of Science* 47 (1996), 415-440.

[7] For Carnap's first turn away from foundationalism, see Rudolf Carnap, "On Protocol Sentences" (trans. Richard Creath and Richard Nollan), *Noûs* 21 (1987 [1932]): 457-70, as well as the commentary by Richard Creath: "Some Remarks on 'Protocol Sentences'," *Noûs* 21 (1987): 471-75.

[8] Karl Popper, *The Logic of Scientific Discovery* (New York: Basic Books, 1959 [1934], 111 (emphasis removed).

[9] The term "basic statements" is Karl Popper's, while "protocol statements" is Rudolf Carnap's and Otto Neurath's. See Creath, "Some Remarks."

[10] Popper, *Logic*, 84, 100-103.

[11] Popper, *Logic*, 102-3.

would indicate not a weakness in his view but a "failure of language as a means of universal communication."[12]

In that regard, the new logical positivism retained its earlier belief that at least some observation is basic. This shows a certain discrepancy in its position: while claiming that, in a non-foundational world, all beliefs and judgments are susceptible to being tested and rejected, logical positivists continued to believe that some experiences are so basic that it is unimaginable that they be shown wrong. This suggested that, from their point of view, some experience was simply not affected by the rest of scientific observers' webs of belief. Thus logical positivists espoused meaning holism: they thought the meaning of one term is affected by the meanings of connected terms, and so that theories must be tested together, as webs of belief. However, logical positivists drew limited implications from this meaning holism, thinking that some beliefs are not affected by changes in other beliefs.[13]

The notion of basic statements shaped logical positivism's account of non-foundational justification. According to the tradition, a scientific theory is justified insofar as its axioms, also known as "postulates" or "primitive propositions," yield observational statements ($OS_{D,}$ or deduced observational statements) that correspond to the actual observational statements made by the scientific community (OS_A, or actual observational statements). The actual observational statements (OS_A) form the subject matter of a scientific theory and the theory has to explicate this subject matter merely by axioms and deductions by "purely logical or mathematical transformations."[14] On this logical positivist picture, scientific theories depend on the agreement in the judgments of all normal observers: without such agreement, scientific theory would lack the subject matter. It is crucial for the understanding of Rawls's early ethical theory that, without the foundational bedrock, overlap in the judgments of scientific observers became critical for justifying a scientific theory.

In terms of ethics, the most problematic part of logical positivism was its criterion of meaning, or its view of what counted as a meaningful statement. Using this so-called criterion of meaning, also known as the "criterion of demarcation" or "principle of verification," logical positivists classified only two types of propositions as meaningful: analytic truths, or propositions true by definition, and synthetic truths, of propositions true in virtue of being verified by experience. [15] The remaining kinds of propositions were "pseudo," or meaningless, propositions.[16] Which propositions counted as meaningful depended on the interpretation of "experience," but it was a typical trope from both foundationalist and nonfoundationalist logical positivists that metaphysics was meaningless since its truths were neither tautological nor

---

[12] Popper, *Logic*, 104.
[13] For a more holistic accounts of experience see Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).
[14] Popper, *Logic*, 71.
[15] "Criterion of demarcation" is Popper's term. See his *Logic*, 40; "Principle of verification" is Feigl and Blumberg's term. See Albert E. Blumberg and Herbert Feigl, "Logical Positivism." *The Journal of Philosophy* 28 (1931): 288. For the content of the criterion of meaning, see Rudolf Carnap, "On the Character of Philosophic Problems," *Philosophy of Science* 1 (1934), 11-12; Ayer, *Language, Truth, and Logic*, 76-77.
[16] Rudolf Carnap, "The Elimination of Metaphysics Through Logical Analysis," transl. Arthur Pap. In *Logical Positivism*, ed. A.J. Ayer (Westport, CT: Greenwood Press, 1959), 61.

verifiable by experience.[17] Similar lines of argument, as we will see, were also leveled against ethics, and innovators within the logical positivist tradition, such as John Rawls, would have to reject this restricting criterion of meaning.

These commitments set the main tasks for logical positivists. Logical positivists concluded that the task of philosophy was the clarification of the conceptual system of knowledge, with two goals in mind: sifting out knowledge from pseudo-knowledge by exposing questions and propositions that are pseudo-questions and pseudo-propositions, and setting knowledge on epistemologically proper grounds by showing that all meaningful empirical propositions are reducible to basic statements.[18]

Combining all these commitments, logical positivism emerged as a very formalistic movement in its neglect of intentional meaning or the content of the assertion. For the most part, logical positivists were not interested in analyzing ethical utterances in search of patterns in their intentional meanings. They lacked this interest for different reasons. The early foundational positivists thought that such a study was impossible. Carnap, for instance, constructed the conceptual system for science drawing only on the structural relations between the given experiences, explicitly disregarding the contents of these experiences because these contents were "incomparable."[19] The later, non-foundational, positivists kept the tradition's disinterest in intentional meaning: Carnap, for instance, aimed at elaborating the logical syntax of language, or what he called the "formal theory of the linguistic forms of that language."[20] He explicitly stated that the theory was to make no reference "either to the meaning of the symbols (for example, the words) or to the sense of the expressions (e.g. the sentences)…."[21] This disinterest in the intentional meaning of ethical terms, when adopted in ethical and political philosophy, would lead to very structural analyses of terms like "good" or "justice." As we will see, it would also affect the character of Rawls's initial ethical theory.

When Rawls studied logical positivist works, he encountered this tradition in its later, non-foundational form. In his earliest available essay, "A Brief Inquiry into the Nature and Function of Ethical Theory" (1946), Rawls referred to such works as Neurath's *Foundations of the Social Sciences*, Carnap's *Philosophy and Logical Syntax* and *Introduction to Semantics*, Hans Reichenbach's *Experience and Prediction*, and Popper's *Logic of Scientific Inquiry*.[22] By that time logical positivism had already opened one of the more interesting periods in analytic philosophy, defending knowledge on non-foundational grounds. The opening of these questions broadened logical positivism's conception of experience to the detriment of the critical edge of

---

[17] See, for example, Carnap, "Elimination of Metaphysics" (entire) and Popper, *Logic*, 35-40.
[18] Carnap's *Logical Structure of the World* is representative of this second type of work, but see also publications from members of the International Encyclopedia of Unified Science Movement, such as Philipp Frank, *Foundations of Physics*. International Encyclopedia of Unified Science I:7 (Chicago: University of Chicago Press, 1946) and Otto Neurath, *Foundations of the Social Sciences*. International Encyclopedia of Unified Science II:1. Chicago: University of Chicago Press, 1944.
[19] Carnap, *Logical Structure of the World*, 106-107, 21, 128.
[20] Rudolf Carnap, *The Logical Syntax of Language* (London: Kegan Paul, 1937), 1.
[21] Carnap, *Logical Syntax*, 1.
[22] Rudolf Carnap, *Philosophy and Logical Syntax* (London: K.Paul, Trench, Trubner & Co., 1935); Rudolf Carnap, *Introduction to Semantics and Formalization of Logic* (Cambridge, MA: Harvard University Press, 1942); Hans Reichenbach, *Experience and Prediction* (Chicago: The University of Chicago Press, 1938).

its analytic-synthetic distinction. As the notion of experience broadened, logical positivism's exclusion of metaphysics and ethics as meaningful disciplines became questionable even among members of the movement. Consequently, when Rawls came under the influence of this tradition, the doors to ethical – although not metaphysical – inquiries were being opened from the logical positivist perspective. He would bring this non-foundationalist, deductive and formalistic conception of philosophy to ethics, thereby also expanding the possibilities of logical positivism in ethics.

Logical Positivism in Ethics

This evolution of the logical positivist movement is seen in its changing positions in ethics. While in its earliest years the tradition was dominated by naturalism and emotivism, by the time Rawls was writing his first graduate essays at Princeton, a new way of thinking about ethics was taking shape. As it modeled ethical inquiry on the logical positivist conception of scientific theory, I will call it the "scientific ethics" position. It is important that, while Rawls drew on the logical positivist tradition to form his own conception of philosophy, he rejected the two dominant logical positivist currents in ethics – including that of his teacher and dissertation adviser, Walter Terrence Stace. That is why his position is better described as positivist, not logical positivist.

Naturalism was the most fruitful ethical current of early logical positivism. It held onto a version of the analytic-synthetic distinction, but attempted to show that ethical statements are reducible to empirical statements and are therefore meaningful.[23] Thus, according to naturalism, ethical statements were hidden empirical statements. Walter Stace pursued precisely this path. Offering a "radical empiricist account of morals," he argued that meaningful statements had to refer, or purport to refer to, "something of a kind whose elements are at least theoretically capable of being directly experienced."[24] He thought that these directly experienced elements were human desires or ends, and attempted to show that ethical statements can be translated into imperatives that lead to these ends "without loss of intended content."[25] For instance, Stace argued that the categorical imperative 'you ought not to overeat' is translatable into the hypothetical imperative 'you ought not to overeat if you wish to retain your health,' which is in turn translatable into an empirically verifiable statement, 'abstention from over-eating is one of

[23] It would be more appropriate to call this current "positivist reductionism," since the term naturalism, as it was originally used, is too broad. Its coiner G.E. Moore understood naturalism very broadly, as any tradition that translates ethical statements into statements about natural qualities, such as desires or states of pleasure. See Moore's *Principia Ethica* (Cambridge: Cambridge University Press, 1968 [1903]), 12-21. Intellectual historians have since used this terms to include not only logical positivist, but also pragmatist and, in principle, idealist reductionism (see William Frankena, "Ethical Theory," in Schlatter, R. (ed.) *Humanistic Scholarship in America: Philosophy*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1964, 364). Using the narrower term "positivist reductionism" would place emphasis on the current's origins in logical positivism. However, in the 1960s Rawls would refer to his own position as naturalist and would try to find ways in which his position does not commit Moore's naturalistic fallacy. To use the terms consistently throughout the dissertation, I keep the term "naturalism," specifying its meaning in the context as far as possible.

[24] Walter Terrence Stace, *The Concept of Morals* (New York: the Macmillan Company, 1939), 17-18.

[25] Stace, *Concept of Morals*, 21.

the means to health.'[26] From thereon, Stace's strategy was twofold: to show that there were universal human ends, and that morality was not only the best, but also the only means to these ends.[27] Since both of these claims were empirically verifiable, the proposed system of ethical statements was meaningful, objective, and, Stace argued, true.

But in ethics logical positivism is best known for its second tradition: emotivism. Emotivism arose to some extent in opposition to naturalism's reductionism, as its two main representatives, Alfred Jules Ayer [1910-1989] and Charles Stevenson [1908-1979], argued that the proposed translation of ethical to empirical statements failed. Ethical statements were not statements about the speaker's state of mind: "this cake is good" did not mean "I like this cake."[28] According to emotivists, such and similar translations left out the distinctively ethical – emotive – meaning.[29] In line with the formalism of the logical positivist tradition, they understood emotive meaning in functional terms: it was thought to be the purpose, or use of ethical statements. According to both Ayer and Stevenson, this purpose was expressing emotions and inducing them in others: "the power that the word acquires … to evoke or directly express attitudes, as distinct from describing or designating."[30] Thus, statements such as "x is good" could be translated into "I approve of x" and "I want you to do so as well."[31] Emphasizing the functional meaning of ethical statements, emotivists understood these statements as imperatives or commands. But, as such statements did not fall into analytic or synthetic categories, they were not in the literal sense significant, but rather "expressions of emotion which can be neither true nor false."[32] This position was shared by most notable logical positivists, including Carnap, who argued that ethical sentences are "pseudo-sentences … [and] have no logical content, but are only expressions of feeling which in their turn stimulate feelings and volitional tendencies on the part of the hearer."[33] Thinking that ethical statements cannot in the literal sense be true or false, most logical positivists thought that the task of a philosopher in ethics was brief: apart from categorizing ethical statements as commands and imperatives, it consisted "simply in saying that ethical concepts are pseudo-concepts and therefore unanalyzable."[34]

Both naturalism and emotivism centered their ethical positions on the analytic-synthetic distinction. By the early 1940s, as the boundaries of logical positivism were becoming more fluid, a new type of argument about ethics emerged in the tradition. In a novel fashion, it skirted the analytic-synthetic distinction and tried to make ethics into an empirical inquiry instead. This new type of argument was developed by Curt John Ducasse [1881-1969], but except for shaping

---

[26] Stace, *Concept of Morals*, 21-22
[27] Stace, *Concept of Morals*, 262-294.
[28] Ayer, *Language, Truth and Logic*, 107.
[29] Ayer, *Language, Truth and Logic*, 105. Charles L. Stevenson was more nuanced in his criticism, stating that "an ethical judgment *can* be true or false, but to point out that its descriptive truth may be insufficient to support its emotive repercussions." See his *Ethics and Language* (New Haven, CT: Yale University Press, 1944), 267.
[30] Stevenson, *Ethics and Language*, 33.
[31] Stevenson, *Ethics and Language*, 24-6. See also Ayer, *Language, Truth, and Logic*, 108.
[32] Ayer, *Language, Truth, and Logic*, 103, 108. Stevenson, *Ethics and Language*, 30.
[33] Carnap, *Logical Syntax of Language*, 278. See also Carnap, "Elimination of Metaphysics," 77. For Feigl and Blumberg's claim that ethics "as 'normative' science is impossible," See Blumberg and Feigl, 293. For Neurath's argument that ethical statements are not translatable into the language of science because they are metaphysical terms, see Neurath, *Foundations of the Social Sciences*, 11.
[34] Ayer, *Language, Truth, and Logic*, 112.

the thought of its most famous follower, John Rawls, this view did not draw followers.[35] Its core idea to model ethical inquiry on scientific grounds was simple and worn in the broader modernist movement of which logical positivism was part.[36] French economist Jacques Rueff [1896-1978], whose *From the Physical to the Social Sciences* [1929] inspired Ducasse, had also proposed introducing the method of the physical science to ethics and constructing "a system of initial propositions, axioms and definitions capable of serving as premises to reasoning."[37] Nonetheless, fashioning ethics as a scientific inquiry was a new and unusual proposal among logical positivists who generally assumed that ethical judgments were fundamentally different from scientific judgments.

Ducasse set out to model ethics after the logical positivist conception of science while ignoring the analytic-synthetic distinction. To do so, he had to show that it had a subject matter of its own and that this subject matter was susceptible to being treated by the scientific method. Ducasse's formulation of these two notions showed logical positivism's influence. He first defined the "primitive subject matter of ethics," or those facts that are "beyond question" and "about which … questions [are] asked by ethical science."[38] This subject matter consisted of ethical judgments and included both particular judgments such as "This is wrong" and empirical generalizations such as "stealing is wrong."[39] Ducasse set two conditions to these ethical judgments: they were to be "most confident" and spontaneous, or made without deliberate application of any ethical theory.[40]

Ducasse's conception of scientific method was typical of logical positivism. He thought that the aim of a theorist both in the natural sciences and in ethics was to formulate axioms or "premises from which could have been deduced … empirically discovered generalizations [such as 'stealing is wrong'] and … others empirically discoverable."[41] These deduced generalizations G_D were subsequently to be tested by actual ethical judgments J_A, and the theory was to be considered justified insofar as the deduced generalizations G_D matched, or predicted, actual ethical judgments J_A.[42] Ducasse did not specify the extent to which the axioms had to predict actual ethical judgments, but, given that he included only confident judgments in the subject matter of ethical theory, he must have thought that the axioms had to predict actual ethical judgments with complete or nearly complete accuracy.

Albeit simple and in the broader philosophical landscape worn, Ducasse's theory opened up the possibility of a truly normative logical positivist ethics. If ethical inquiry was successful, it would result in ethical principles that all persons acknowledged as true of at least their most

---

[35] Curt John Ducasse, "The Nature of Function of Theory in Ethics," *Ethics* 51 (1940): 22-37; Curt John Ducasse, *Philosophy as a Science: Its Matter and Method* (New York: O. Piest, 1941).
[36] For the relationship between logical positivism and modernism, see Peter Galison, "Aufbau/Bauhaus: Logical Positivism and Architectural Modernism", *Critical Inquiry* 16 (1990), 709-52 and Peter Galison, "Constructing Modernism: The Cultural Location of the *Aufbau*", in Ronald N. Giere and Alan W. Richardson, eds., *Origins of Logical Empiricism* (Minneapolis: University of Minnesota Press, 1997), 17-44.
[37] Jacques Rueff, *From the Physical to the Social Sciences*, trans. by Herman Green. (Baltimore: John Hopkins University Press, 1929), 65.
[38] Ducasse, "Theory in Ethics," 28-29.
[39] Ducasse, "Theory in Ethics," 28-29. Cf. Ducasse, *Philosophy as a Science*, 74.
[40] Ducasse, "Theory in Ethics," 29.
[41] Ducasse, "Theory in Ethics," 30.
[42] Ducasse, "Theory in Ethics," 29, 32.

confident judgments. Assuming human wish to be consistent, these principles would become a powerful normative force. On the individual level, the principles could reveal inconsistency between the judgments and the principles, thereby also informing us "of the alterations to be made in [these divergent judgments]."[43] Given the assumption that all observers agree in their judgments, ethical inquiry was meant to function in exactly the same manner on the social level. The principles were meant to function as an internal critique, informing us of the alterations to be made in order to resolve inconsistencies among these judgments – in this case disagreements between different persons.[44] To fulfill this normative promise of the scientific ethics position, one had to show that ethical judgments of all persons converged sufficiently to permit the formulation of ethical principles.

Ducasse's proposal, while promising, was unfinished. He did not attempt to show that all persons would actually agree in their ethical judgments, and, most importantly, he did not explain why he thought all persons would agree. As his proposal stood, the emotivist argument that ethical judgments were simply expressions of emotion was skirted but not rejected. Rawls, impressed by the scientific edifice of which ethics was thought to be capable, would undertake to show that all persons would agree in their ethical judgments. Taking on this task, he would also be forced to engage the difficult questions which Ducasse ignored.

A "Physicalist" Approach to Ethical Theory

Rawls's positivism evolved even during the initial five years covered in this chapter; his position in 1946 differed from that which he held between 1947 and 1951. In 1946, Rawls's thinking was noticeably influenced by Popper and Ducasse: he modeled ethical inquiry on scientific theory and took the first steps in defining its subject matter. The result was a very formal ethical inquiry, interested in the function of ethical utterances and not their intentional meaning. From 1947 onward this picture of ethical inquiry began to change. For the most part, Rawls continued to develop this scientific conception of ethics, elaborating the conception of objectivity and taking the first steps toward showing that all reasonable persons agree in their ethical judgments. The newly developed notions of the reasonable man and rational judgment are examples of such developments. However, 1947 marked the first noticeable influence of linguistic philosophy: Rawls introduced a new notion of justification that relied on our shared sense of "absurdity." Between 1947 and 1950, the notion of absurdity had little systematic effect on Rawls's overall theory and only influenced his conception of ethical principles. But the change was significant, as it was Rawls's first of the many moves away from the strict positivism of his initial years.

The shaping influence of logical positivism is reflected in Rawls's graduate school writings. "A Brief Inquiry into the Nature and Function of Ethical Theory" (1946), his earliest and only surviving essay from that year, was typically non-foundational, relying for the justification of its views on the consistency between the axioms of a theory and the ethical

[43] Ducasse, "Theory in Ethics," 35.
[44] Ducasse, "Theory in Ethics," 35-36.

judgments it was meant to explain.[45] It contained a familiar conception of deductive scientific theory. And it was markedly formalistic in its explicit lack of interest in the intentional meaning of ethical utterances. Its main goal was to show that all reasonable persons agree in their ethical judgments. But this agreement was agreement in the uses of ethical words such as 'right,' not on the criteria of right actions.

It is important that Rawls saw himself as part of the logical positivist movement. In 1946, he described his theory as "physicalist in the same sense as this term was understood by the Vienna Circle ([in] essays in Erkenntnis by Carnap [and] Neurath)."[46] Noting the novelty of his approach within the tradition, he portrayed his theory as an extension of the physicalist theory to ethics:

> It is the business of philosophers to begin an inquiry, to break the ground, to so formulate and clarify the domain of investigation that it can become an exact science. Philosophers have already performed this duty for physics, astronomy, psychology and other sciences. The task remains to be done for ethics, and this essay is such an attempt.[47]

Extending the "physicalist" theory, Rawls drew mostly on Ducasse, whose essay he praised as "excellent throughout on many points discussed here."[48] Even the title of his essay is an acknowledgement of Ducasse's influence.[49] Rawls thought that ethics diverged from other scientific disciplines only in its subject matter: "the technique of theory construction is the same in ethics as it is in physics. The only difference concerns subject matter."[50] The subject matter "peculiar to ethics," accordingly to Rawls, "is the facts of ethical judgment."[51]

Rawls's conception of scientific theory was typical of logical positivists: he thought that the scientist aimed at elaborating axioms from which the predicted particular ethical judgments (J_D) would be deduced and then tested against actual ethical judgments (J_A). Quoting Rueff, Rawls wrote that philosopher's task in ethics was to formulate principles which can "predict [these] judgments":

> The following quotation from Rueff, cited by Ducasse …, expresses perfectly the viewpoint here presented in slightly different words: the task of ethical theory is to '… enunciate a system of initial propositions, axioms, and definitions which, when fed into the reasoning machine, will produce theorems coinciding with the rules of practical morals.'[52]

---

[45] John Rawls, 'A Brief Inquiry into the Nature and Function of Ethical Theory', The Papers of John Rawls, Harvard University Archives, HUM 48 Box 7, Folder 3 (1946).
[46] Rawls, "Nature of Ethical Theory," 9. Rawls's reference is to *Erkenntnis* 3:2 (1932/1933), which contains Neurath's "Protocol Statements" (pp.204-214) and Carnap's "On Protocol Statements" (pp.215-228). The underlining is Rawls's, as in the rest of the dissertation.
[47] Rawls. "Nature of Ethical Theory," 61.
[48] Rawls, "Nature of Ethical Theory," 9.
[49] Rawls only added "A Brief Inquiry to" to Ducasse's "The Nature and Function of Theory in Ethics."
[50] Rawls, "Nature of Ethical Theory," 7-8.
[51] Rawls, "Nature of Ethical Theory," 7.
[52] Rawls, "Nature of Ethical Theory," 53-54, 9.

In "Nature of Ethical Theory," Rawls took it as his task to construct and justify just such a theory:

> We propose to construct a theory, to make deductions from it, and to test these deductions against the subject matter of ethical theory, namely, the actual moral judgments made by the class of people whose judgments constitute the reference of the theory.[53]

Logical positivist themes that shaped the skeleton of Rawls's conception of ethical theory also influenced its central features. The key notion of ethical judgments was designed with Popper's basic statements in mind.[54] Modeling ethical inquiry after the image of empirical theories, Rawls fashioned ethical judgments as the data against which theories are tested:

> The physical sciences have as their subject matter certain processes which might be termed 'thing' processes. And every physical theory is testable in that it denies that certain 'thing' processes ever occur. Now ethical theory is essentially the same. Its subject matter, however, is not a 'thing' process, but a 'word' process, and every ethical theory is testable in that it denies that certain specified 'word' processes ever occur.[55]

Rawls's understanding of what was counted as "testable" was close to being behavioralistic: it stressed the actual utterance of the ethical statement and was uninterested in the meaning attributed to the expressed statement. The key requirement for ethical judgments was that they be uttered: "we discover what a person means to assert by observing his subsequent behavior. And so it is in ethics. To determine what people mean to assert by ethical statements, we observe how they use the word, and how they act within the 'sign-context'."[56] Although Rawls's explanation suggested that the philosopher was interested in "what people mean to assert," that is, reasons for which normal observers judged something just or unjust, his examples indicated otherwise: ethical judgments were of the type "this act is right (wrong)."[57] Reasons for which the judgments were made were not relevant for ethical theory. Indeed, he argued that these "individual mental contents" were impossible to observe and hence unsuitable for an empirical theory: "statements about mental contents cannot be asserted and supported by any adequate technique."[58]

The second requirement for ethical judgments was that they express our "deep seated convictions" and "deepest feelings," for which reason we were also "most certain of" these judgments.[59] Although Rawls may have had other explanations for our certainty in these judgments, in "Nature of Ethical Theory" he modeled these most certain judgments after basic statements. He thought that ethical judgments fitting to serve as the subject matter of ethical

---

[53] Rawls, "Nature of Ethical Theory," 29.
[54] Rawls, "Nature of Ethical Theory," 20.
[55] Rawls, "Nature of Ethical Theory," 7.
[56] Rawls, "Nature of Ethical Theory," 19.
[57] Rawls, "Nature of Ethical Theory," 9.
[58] Rawls, "Nature of Ethical Theory," 20-21. In his argument, Rawls referred to Charles Morris, *Signs, Language and Behavior* (New York: Prentice Hall, 1946), 28-29.
[59] Rawls, "Nature of Ethical Theory," 53-54.

theory were inductivist generalizations of the judgments made by different "normal observers."[60] Such judgments had to be rendered by a sufficient number of normal observers. Again, Rawls's analogy was to reasoning in science: "'There is a tree' is an assertion to the effect that all, or most all, of a class of normal observers will assert, under specified conditions, 'I see a tree at such and such etc'."[61] Rawls compared the more general statements – which, following Popper, he called basic statements or a "basissatze" – to probability statements: "A 'Basissatz' is really a probability statement claiming a very high, though not exactly specified, frequency of 'Erlebnissatze' [perceptual judgments] of the form 'I see such and such at such and such' etc."[62]

By analogy, the idea was to show that a sufficiently high proportion of "normal observers" agreed in judgments such as "x is just." Achieving this would provide the subject matter for a scientific ethical theory. It was an ambitious task, all the more because Rawls, like the rest of logical positivists, was universalist in the scope of his theory, including in the group of "normal observers" "all animals which are capable of using, understanding, and acting on such word processes as 'this is right (wrong)' etc in whatever word-language they may be uttered."[63] Like the positivists who inspired his work, Rawls took the existence of this agreement for granted. Showing that this agreement exists would become Rawls's main goal between 1946 and 1951. He started this task in 1946, by proposing his own ethical theory, "Imperative Utilitarianism."[64]

Imperative Utilitarianism

Imperative utilitarianism was unexpectedly formalistic: Rawls was interested not in the content of ethical judgments but in their semantic meaning, or the function which the terms 'right' and 'wrong' have in human life. Most notably, in 1946 Rawls was interested not in what actions or institutional arrangements ethical concepts require but in purposes for which such concepts are used in human life. Not less strikingly given his position in *A Theory of Justice*, Rawls's first ethical theory was self-avowedly utilitarian – although, as we will see, this self-description needs crucial qualifications. These significant differences between Rawls's early positivist theory and the one he would propose in *A Theory of Justice* shows that his thought would undergo important changes in the twenty five years that separate these works.

First, then, Rawls's theory was formalistic: it was interested in the semantic meaning of 'right' and 'wrong.'[65] This formalism is best illustrated in his treatment of the later Ludwig Wittgenstein's argument that the search for regularities in the meanings of words was unlikely to succeed because, as the word was placed in a different conceptual background in different instances of its use, its meaning changed accordingly.[66] Rawls acknowledged the force of this

[60] Popper himself rejected induction as a way of testing the more general statements of science and axioms of a scientific theory because he thought that David Hume's objections to induction were insurmountable.
[61] Rawls, "Nature of Ethical Theory," 20.
[62] Rawls, "Nature of Ethical Theory," 20.
[63] Rawls, "Nature of Ethical Theory," 9.
[64] Rawls, "Nature of Ethical Theory," 48.
[65] Rawls, "Nature of Ethical Theory," 29.
[66] See, for example, Ludwig Wittgenstein, *The Blue and Brown Books* (New York, 1965), 17-8.

objection, agreeing that, in phrases 'a good race horse,' 'a good work horse,' and 'a good horse for children to ride,' "'good' means something different according to the context, according to the 'thing' to which it is applied."[67] However, he thought that Wittgenstein's objection was irrelevant because ethical theory was interested not in the criteria for ethical terms, but in what he called their "semantic" meaning: the use of the ethical expression to do something else, or "a certain operation of selection in terms of the characteristics of the things referred to" that the word "means to perform."[68] Thus, 'good,' while it did not have a common intentional meaning, still had a common semantic meaning – to direct the interlocutor's attention to qualities that make particular things good:

> In applying the word 'good' to a thing in the attributive sense we are directing the hearer to perform an operation of selection on the qualities of the subject of the attribution according to certain definite principles such as the principles of successful fulfillment of purpose involved in the usual use of the thing.[69]

It is important that Rawls's self-description as a "utilitarian" be understood in the context of seeking to explain the semantic and not the intentional meaning of ethical words. In "Nature of Ethical Theory" Rawls concluded that 'right' and 'wrong' are used to encourage rare actions which the speaker thinks "will lead to the greatest amount of good" (in the case of 'right') and to discourage frequent actions which the speaker thinks diminish the amount of good (in the case of 'wrong').[70] As such, the use of 'right' and 'wrong' seemed utilitarian to Rawls. Like utilitarians, he wanted to show that, despite the different virtues encouraged by different societies, there was a common function to the use of ethical words. However, Rawls was not a utilitarian in a substantive way: he did not identify the good with pleasure or satisfaction of a desire. In "Nature of Ethical Theory," he only wanted to explain the cultural variation in ethical judgments. He did so by noting that the "contextual occurrence" of different activities, or the "frequency with which [they] are met with in social life," differed among societies. Thus a nation surrounded by hostile neighbors and frequently engaged in war will praise "the virtues of the soldier," such as bravery, obedience, endurance, devotion, and loyalty.[71] On the contrary, a nation that spends most of its efforts on commerce will praise the virtues of industriousness, thrift, cunning in dealing with foreigners, and the like.[72] In this way, contextual occurrence explained the change in appraisals in the same society over time. This explanation itself was formalistic: Rawls was interested not in the reasons for which different societies made their decisions but in contextual factors, such as proximity to warlike neighbors.

Once qualified as a theory of the nature of ethical statements, Rawls's utilitarianism becomes intelligible as an extension of his logical positivism. In fact, it also shows that, despite his rejection of emotivism, Rawls was nonetheless influenced by its formalistic analysis of ethical terms. Emotivists claimed that ethical utterances, insofar as they are meaningful, should be understood as imperatives or commands meant to incite appropriate feelings or induce

---

[67] Rawls, "Nature of Ethical Theory," 55-56.
[68] Rawls, "Nature of Ethical Theory," 56. Rawls also describes semantic meaning as "linguistic function" of a word. Ibid, 25.
[69] Rawls, "Nature of Ethical Theory," 56.
[70] Rawls, "Nature of Ethical Theory," 29-30.
[71] Rawls, "Nature of Ethical Theory," 43.
[72] Rawls, "Nature of Ethical Theory," 43.

appropriate behavior. It is not a coincidence that Rawls called his theory "imperative utilitarianism": his analysis of ethical terms was similarly oriented to the purposes with which ethical words are used and not the criteria associated with these words. Rawls too interpreted ethical terms as "imperatives, functioning to increase or decrease, as the case may be, [the frequency of the mentioned actions]."[73] He quibbled with Ayer, claiming that ethical statements were in fact like imperatives, similar to them in some respects but different in others, but he never detailed the ways in which ethical statements differed from imperatives.[74] Thus Rawls's earliest ethical theory not only drew on key logical positivist themes, but also bore important commonalities with emotivism, the standard logical positivist expression in ethics.

Political Theory, Political Practice

Rawls's achievements in political theory and philosophy are associated not only with his philosophical vision, but also with his political vision. The resurgence of political thought, to which Rawls contributed, came about on the shoulders of new philosophical visions. And while philosophy does not straightforwardly imply political views, we nonetheless want to ask how Rawls himself employed his philosophical framework to recommend particular political practices. In 1946, Rawls certainly intended imperative utilitarianism to guide our political practice, but he did not specify any particular ethical or political vision which it implied.

Rawls thought that political theory would guide political practice by serving as an immanent critique. In this argument he followed Ducasse's reasoning. Assuming the truth of the key positivist hypothesis that a "very high, though not exactly specified" proportion of "normal observers" agree in their judgments – the truth which Rawls claimed to have exhibited in imperative utilitarianism – ethical principles would represent a stable point in the swamp from which other arguments would follow. In case of disagreement, one could use these principles to draw deductions and see what judgments they require in particular cases:

> We require a theory whose predictions correspond to our 'deepest intuitions.' Once we have such a theory, it can function as a mediator in cases of conflict. We can say to the disputants that the theory in question explains what their moral judgments <u>really</u> are. If our theory is adequate to forecast their 'deepest feelings' they will be convinced, and assuming they wish to be consistent, they will agree to resolve the conflict by applying the moral imperative according to the dictates of the theory.[75]

Rawls did not go into the particulars of how the principles would function to recommend any particular actions, but we can make some conjectures. For instance, it is possible to turn contextual factors into reasons when deliberating about practical politics. Thus, if Rawls's theory claimed that a society most highly values valor because it is surrounded by warlike neighbors, "being surrounded by warlike neighbors" could become a reason in deliberation. While by itself this reason would not be sufficient to lead all reasonable persons to agree on a course of action, it

[73] Rawls, "Nature of Ethical Theory," 30
[74] Rawls, "Nature of Ethical Theory," 25.
[75] Rawls, "Nature of Theory," 53-4; cf. Ducasse, "Theory in Ethics," 36.

may have significant force, especially if, prior to the consideration of Rawls's theory, this reason was unduly neglected. Admittedly, then, the exact nature of the connection between theory and practice is a matter of speculation, and it is evident that in this regard Rawls's early ethical theory differed sharply from that of *A Theory of Justice*. Nonetheless, the guiding idea that ethical theory should guide ethical practice was already there.

In sum, Rawls's approach to philosophy in 1946 was shaped by logical positivist themes. In the context of the tradition's typical positions on ethics, his approach was novel: it was centered on the claim that ethical theory, like all scientific theories, is an empirical theory. Defining the subject matter of ethical theory and applying to it the picture of scientific theory, Rawls followed key logical positivists such as Popper, Carnap, and Neurath. Adopting this picture of scientific inquiry, Rawls took upon other features of their thought, in particular their non-foundationalism. Nonetheless, there remained a question of the extent to which positivism could be turned into an ethical theory because, as positivist as Rawls's theory was, it did not seem to appreciate sufficiently the core emotivist objection that ethical judgments are simply expressions of emotion. Rawls thought that the emotivist conclusion rested only on the failure to find commonality in our ethical judgments.[76] Thinking that this impossibility of finding agreement in our judgments was yet to be shown, he offered his own theory as an example that such agreement was indeed possible. He did not engage emotivists' broader point that ethical judgments cannot be true or false, reasonable or unreasonable – in brief, objective – because they are expressions of emotion. Some relation between emotion and human agreement had to be drawn to avoid the impression that Rawls was building a scientific theory of ethics despite the emotivist objection. Between 1947 and 1951, Rawls would become aware of the need to respond to this broader emotivist claim. In 1947, he would do so by elaborating a conception of objectivity that still skirted this objection, but did so explicitly, explaining why ethics did not need to address it.

Scientific Conception of Objectivity

In 1947, Rawls left Princeton for a year at Cornell. Likely, he knew that his former teacher Norman Malcolm was to join the faculty at Ithaca that fall.[77] If he did, he decided not to take any seminars with Malcolm and engaged in private conversations with him instead. Rawls's acquaintance with Malcolm dated back to Malcolm's 1942 class on social philosophy and resumed with his 1946 seminar on the theory of knowledge.[78] It was probably Malcolm who introduced Rawls to Wittgenstein's *Blue Book*, which Rawls listed in the bibliography of "Nature

[76] Rawls, "Nature of Ethical Theory," 6-7.

[77] Rawls took courses with the historian of science Henry Guerlac and philosopher Max Black, who joined Cornell the same semester as Malcolm, Fall 1947, although there is no evidence to suggest that Rawls knew of their arrival to Cornell in 1946. He also engaged in individual conversations with philosopher Arthur Edward Murphy, who had been at Cornell since 1945. While it is possible that Rawls decided to spend the year at Cornell in order to study with Murphy, I find no intellectual affinities between the two thinkers.

[78] John Rawls, "Letter to Robert Audi" (1978), 1. I still have to confirm the location of this letter but, to the best of my knowledge, it is in the alphabetically arranged correspondence folder in John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 43, Folder 1.

of Ethical Theory" [1946].[79] But Malcolm's influence started to show only after the conversations at Cornell. These conversations would give Rawls new concepts for thinking about philosophy and within a decade would lead him to reject large parts of his positivist framework. Referring perhaps to this intellectual development, in 1978 Rawls called Malcolm his "most influential teacher": one who, "in more ways no doubt than I can say, … has left his mark on the way I think of philosophy."[80]

This new way of thinking was linguistic philosophy. Malcolm was a friend and student of Wittgenstein; the two frequently conversed at Cambridge throughout the academic year 1946-47. Taken by the themes of this new conception of philosophy, he would become one of the first thinkers to bring it to the U.S.[81] Because Rawls adopted only the surface arguments of linguistic philosophy between 1947 and 1951, I will leave the discussion of linguistic philosophy for the next chapter. Here it suffices to say that the tradition's commitments and tasks stemmed from its key belief that language, like all human practices, is governed by roughly shared rules. In 1947, this key belief was relevant to Rawls in two ways. To determine the rules of various words, linguistic philosophers examined the contexts in which these words were used. Doing so, they paid attention to the intentional meaning of the word, noting its varied conceptual connections in different contexts of its use. This was not the only interest of linguistic philosophers: J.L. Austin's *How to Do Things with Words* shows that they were also interested in the uses of words, or what Rawls called the semantic meaning.[82] Nonetheless, studies of the intentional meanings of concepts such as 'certainty' were typical. This would be significant to Rawls whose "Nature of Ethical Theory" was interested in the uses of 'right' but not in the different intentional meanings of this word.

Between 1947 and 1951, linguistic philosophy most influenced Rawls through its notion of "absurdity" or "nonsense." The notion of nonsense was not original to linguistic philosophers: logical positivists themselves used this concept, claiming that any statement which is neither analytic nor synthetic is nonsensical. Indeed, the rejection of metaphysics and religion as nonsensical was one of the trademarks of logical positivism. However, linguistic philosophers' interpretation of this concept was novel and importantly different from that of logical positivists. They thought that, when words were used against the rules, confusions of thought resulted. Questions or statements stemming from such confusions were thought to be "absurd" or "nonsensical," and persons asking such questions seriously would have to live odd lives – lives unfamiliar to us. Linguistic rules were therefore closely connected to various human practices. Many seemingly deep philosophical questions were thought to rest on confusions of linguistic rules. Like logical positivists, linguistic philosophers saw it as their task to show that these confused questions cannot be meaningfully asked. This was done by bringing to light the absurdity which resulted from asking this question: an argument which was thought to be sufficient to convince everyone, including those who had formulated the nonsensical question. This notion of absurdity – albeit without the key background notion of human practice – would make its way into Rawls's arguments.

---

[79] Wittgenstein, *The Blue and Brown Books*, vii. Rawls, "Nature of Ethical Theory," 64.
[80] Rawls, "Letter to Audi," 1.
[81] For Malcolm's relationship with Wittgenstein, see Malcolm's memoir *Ludwig Wittgenstein: A Memoir* (New York: Oxford University Press, 2001).
[82] John L. Austin, *How to Do Things with Words* (Cambridge, MA: Harvard University Press, 1962).

Between 1947 and 1951, however, Rawls's main preoccupation was formulating the nature of objectivity in ethics. As his 1947 paper "Remarks on Ethics" attests, Rawls's main arguments drew on positivist assumptions: he modeled ethical objectivity on scientific objectivity.[83] He argued that ethical judgments were objective insofar as they satisfied appropriate tests: "the objectivity of science does not depend upon how it is learned, or how it is arrived at, but rather upon its satisfaction of certain tests which we apply to statements once they have been formulated."[84] He assigned three such tests: ethical judgments had to gain agreement of all reasonable persons, this agreement had to be correlated with the occurrence of a relevant objective quality, and any disagreement had to be explained as a result of failure, or "a certain definable illness or peculiarity."[85] Physics provided an example of how such tests are satisfied. In physics there is a general agreement on the use of terms for color, this agreement is correlated with the occurrence of "physical properties," namely, the wave-lengths of the light emitted, and failure to discern a proper color is "paralleled with defects in [one's] organs of vision."[86] To be objective, ethical judgments had to satisfy the three tests in an equivalent way.

Rawls argued that ethical judgments did satisfy these tests. As he proposed to show, all reasonable persons agreed in their ethical judgments. This agreement was correlated with proper motivation to do the right action: "when we study when and where they agree, we find that we can correlate the quality spoken of as the property of a greater and less tendency to do the right "of itself" on the part of human character."[87] And disagreement in ethical judgments was explained by lack of education and disagreement in beliefs about the world: "disagreement can be correlated with inability to learn what right and wrong are, lack of training and education, but most often in common life, with variability of beliefs, ie., people are not even examining the same situation when they believe differently…."[88]

Rawls spent most of his efforts on showing that ethical judgments satisfied the first test: that all persons agreed in their ethical judgments. Between 1947 and 1950, he devised ways to limit the range of ethical judgments which an ethical theory had to explain. As Rawls wrote, "many judgments which we make are not meant to be taken seriously, and many others, we readily admit, do not deserve to be conscientiously considered."[89] To eliminate such judgments, he developed the notions of the "reasonable man" and "rational" or "reasonable" judgments, which would play an important part in *A Theory of Justice*. These notions originated in Rawls's positivist attempt to build an empirical theory of ethics and not, as their long history in law might suggest, in legal philosophy. Rawls was aware of the "reasonable person" and "reasonable" in law, and, undoubtedly, the name of the "reasonable man" was suggested by this use.[90] Yet the content of these notions was shaped by the demands of an empirical ethical theory. In law, the "reasonable person" functions as a standard to evaluate the reasonableness of the defendant's

[83] John Rawls, 'Remarks on Ethics', The Papers of John Rawls, Harvard University Archives, HUM 48 Box 9, Folder 15 (c.1947).
[84] Rawls, "Remarks on Ethics," 8.14.
[85] Rawls, "Remarks on Ethics," 4.19-4.20.
[86] Rawls, "Remarks on Ethics," 4.20.
[87] Rawls, "Remarks on Ethics," 4.20.
[88] Rawls, "Remarks on Ethics," 4.20-4.21. Cf. Rawls, "Grounds of Ethical Knowledge," 276.
[89] Rawls, "Grounds of Ethical Knowledge," 45.
[90] Rawls, "Remarks on Ethics," 5.18. For Rawls's quick dismissal of the American legal realists Karl Llewellyn and Jerome Frank's views that divergence of moral opinions is insurmountable, see Ibid, 5.11.

actions. It draws an intuitive limit to actions that are permissible or required, thereby directing the interlocutor's decision in a determinate direction. But "reasonable person" and "rational judgments" could not play this role in an empirical theory.[91] Non-foundational logical positivist theories relied for their justification on their ability to explain the protocol statements of a universal, or nearly universal, scientific community. Introducing a strong ethical standard in the selection of judgments to be explained would beg the question. The same was true for an empirical ethical theory: it could not introduce a strong ethical standard to tamper with its subject matter. Therefore, "reasonable person" and "rational judgments," which were used to jettison some judgments as unfit for the subject matter of ethical theory, could not be defined by a strong ethical standard "without making the basis of moral principles tautological."[92] By the same reasoning, the weaker the conditions imposed on the notions of the "reasonable man" and "rational judgments," the wider is the range of admissible judgments, and the stronger the support for the claim that ethical judgments are objective.[93] Both in 1947 and in 1950 Rawls's ambition was to make the theory universal.[94] The role of these notions was therefore to restrict the data pool for ethical theory to trustworthy judgments without damaging the empirical basis of the theory.

Rawls defined the notions of "reasonable man" and "rational judgments" with the purposes of empirical theory in mind. A reasonable man had three characteristics: the "ability to understand and to use [the] canons of evidence" by which he "may justify his right to hold an opinion," knowledge of these canons, and willingness to "submit to judgment of these canons to determine what opinions, beliefs, and propositions he shall assert to be true."[95] It is debatable how much normative weight these restrictions actually carried. For example, the ability to judge which evidence is appropriate to ethical questions arguably presupposes some very weighty standards. However, it was not Rawls's intention to define the reasonable man in contestable terms. Indeed, his ambition was to make the theory universal: the reasonable man was said to be any person of reasonable intelligence and moral sensitivity.[96]

The notion of the "rational" or "reasonable" judgments, developed only in Rawls's dissertation, "A Study in the Grounds of Ethical Knowledge" [1950], was also meant to sift off inadmissible judgments.[97] To avoid begging the question, Rawls defined judgments as spontaneous or – tellingly – "empirical," rendered after a "direct and instantaneous" contemplation of the ethical situation and not after a conscious application of some moral rule or theory.[98] They were also to be stable, or "reflecting an enduring disposition to judge in the same way"; impartial, or based on knowledge of relevant interests and not hastily made or favoring unjustly one interest over another; and, finally, certain, or expressing "deep-seated intuitive convictions which remain on reflection."[99] The idea was to exclude judgments based not on their

[91] For Rawls's remarks that ethical theory is an empirical theory, see "Remarks on Ethics," 4.2, 5.1.
[92] Rawls, "Remarks on Ethics," 5.20.
[93] Rawls, "Remarks on Ethics," 5.21.
[94] Rawls, "Remarks on Ethics," 8.12.
[95] Rawls, "Remarks on Ethics," 5.18.
[96] Rawls, "Remarks on Ethics," 5.20.
[97] Rawls, "Grounds of Ethical Knowledge," 60, 66.
[98] Rawls, "Grounds of Ethical Knowledge," 45-49.
[99] Rawls, "Grounds of Ethical Knowledge," 49-52, 52-57, 57-60.

content, but on the way they were rendered. As we will see, these restrictions were weak enough to allow unresolvable disagreements among reasonable men to occur.

1947 introduced another important change: Rawls was now interested in the intentional meaning of ethical terms: he wanted to find agreement in the content of ethical judgments. Rawls now similarly studied "the kinds of actions to which guilt [for instance] is attributed, and the conditions under which such actions occur."[100] His earlier positivist formalism was from then on a thing of the past. In all likelihood this change of interest was a result of linguistic philosophy's influence, as this latter sought to find criteria for the right use of words.

This change made the task of showing agreement of all reasonable persons much more difficult. Moreover, in comparison to his later self, Rawls was strikingly pessimistic about the range of agreement one could expect on ethical issues. He thought that pervasive ethical disagreement was an inevitable result of people's diverging opinions about the world: since the content of beliefs about the world "provides the character of the situation to be examined and judged," ethical judgments will differ as long as they are based on different beliefs about the world.[101] To attain agreement on ethical issues, all reasonable persons had to share these beliefs:

> As long as there are people duped by fantastic magical ideas, and as long as there are those who share a more sophisticated form of it in some kind of Hegelian idealism or its invert, dialectical materialism, and the like, the agreement which would be made known by sharing the truth, is hopelessly covered by distortion.[102]

Given the virtual unattainability of ethical agreement, Rawls's aim to show the objectivity of ethical judgments might have seemed impossible to reach. He thought otherwise, however, offering two types of evidence for the objectivity of ethical judgments. First, he claimed that it was sufficient to show that all reasonable persons agree if they hold the same beliefs about the world: "before instability [of ethical judgments] can be demonstrated it is required that the conflict exists when there is agreement in relation to all relevant beliefs."[103] This response, as odd as it may seem, was intelligible in the contemporary context: it was a sufficient argument against the emotivists who claimed that agreement in belief cannot guarantee agreement in attitude [ethical judgment].[104] Showing that such agreement actually exists was more difficult: it would have required redefining the "reasonable person" in terms of scientific beliefs. This would have been a way to get "at the truth [about the world] and … [adopt] an objective standpoint."[105] This objective standpoint was offered by the presumably "physicalist" scientific temper:

> The firm adoption of the scientific temper of mind will show the underlying convergence of judgment which exists. Only as such a temper spreads throughout the world will this convergence be known.[106]

---

[100] Rawls, "Remarks on Ethics," 2.1.
[101] Rawls, "Remarks on Ethics," 1.1, 7.11. See also Ibid., 8.21.
[102] Rawls, "Remarks on Ethics," 8.21.
[103] Rawls, "Remarks on Ethics," 8.12.
[104] Stevenson, *Ethics and Language*, 31.
[105] Rawls, "Remarks on Ethics," 8.24.
[106] Rawls, "Remarks on Ethics," 8.24.

Rawls did not try to show that, if reasonable persons share the "scientific temper," they also agree in their ethical judgments. It is unclear why he did not go this route; perhaps it is because he saw no agreement even among those who were convinced by the scientific temper, such as logical positivists. More likely, he must have thought that such a move would have endangered the character of his scientific theory by restricting the subject matter significantly. Instead, Rawls chose to provide the second type of evidence for the objectivity of ethical judgments: that, despite the divergence in their beliefs about the world, all reasonable persons agree on at least some ethical judgments. His idea was to appeal to judgments shared by all reasonable persons, "irrespective of [their] other beliefs," and explicate them in terms of "higher-order principles" that were true.[107]

Although this idea calls to mind the notion of the "overlapping consensus" that Rawls would develop in *Political Liberalism* [1993], the two should not be confused.[108] In 1993, Rawls would claim all reasonable persons would agree on reasons that decide controversial political questions.[109] In 1947 and 1950, however, he thought of these higher order principles as background presuppositions to our ethical thinking the rejection of which was unimaginable – absurd – because it would require the rejection of other beliefs and practices we take for granted. Rawls's list of such principles speaks for itself. For example, the first principle required that only acts over which we have control should be judged as indicative of our moral character:

> An act is not to be considered as indicative of the moral worth of the agent's character, unless, in the circumstances under which it was performed, the agent could have done otherwise if he had chosen.[110]

The remaining five principles were similarly general. The second principle, for instance, claimed that the character of a person contemplating an evil action without doing it "is not to be judged as bad as the character of an agent who not only contemplates [an evil action], but does it."[111]

Rawls insisted that these principles help solve some practical ethical problems. For instance, the first principle ruled "as wrong the various forms of political discrimination against racial groups … [which punish] a man or a group for attributes which he or it cannot choose to have or not to have."[112] Despite such practical applications, Rawls acknowledged that "a good number of indeterminate ethical questions will remain."[113] These principles listed "the kinds of actions" to which the term "indicative of moral worth" is attributed, but for the most part they left off the discussions about the kinds of actions that are morally worthy. In comparison to *A Theory of Justice*, which would raise precisely these latter kinds of questions, the practical relevance of the 1947 principles was markedly limited.

Rawls justified these principles by showing that the entire tradition of philosophy affirmed them. He did so in broad strokes, writing that these principles are affirmed by "all

---

[107] Rawls, "Remarks on Ethics," 3.7
[108] Rawls, *Political Liberalism*, 35-40, 133-72.
[109] Rawls, *Political Liberalism*, 133-72, esp. 138.
[110] Rawls, "Grounds of Ethical Knowledge," 110.
[111] Rawls, "Grounds of Ethical Knowledge," 120.
[112] Rawls, "Remarks on Ethics," 5.16.
[113] Rawls, "Remarks on Ethics," 5.17.

ethical theorists as far as I know," "widely recognized," reflect "the moral opinion of men generally," and that there is no one in his knowledge "who has ever denied these principles."[114] When the proposed principles seemed to go against the tradition of ethical thought, he took pains to show that it was a wrong impression.[115] This type of justification was consistent with his conception of ethical inquiry: principles, he thought, were justified by showing that their implications are affirmed by reasonable persons.

Thus in its overall character, "Remarks on Ethics" continued the positivist project started in "Nature of Ethical Theory." Rawls's conception of objectivity was based on the analogy between reasoning in ethics and reasoning in science, and it required that the philosopher show the actual overlap in the judgments of reasonable persons. The "reasonable man" and "rational judgments" were two important steps in that direction, even if initially they did not help to show that reasonable persons agreed on any particular "proposed line of conduct."

Despite this overall logical positivist character of "Remarks on Ethics," Rawls's reasoning in that essay started to show the first signs of linguistic philosophy's influence. In 1947, when he drew on the tradition's notion of "absurdity" without adopting the wider conceptual framework on which it relied, linguistic philosophy's themes were integrated into the overall positivist framework and often came under the guise of analogies between ethics and science.[116] But arguments relying on the notion of "absurdity" were new and already in 1947 they began to change Rawls's thinking. Rawls used the notion of "absurdity" to justify ethical principles by showing that their denial requires living a life so odd that it is nowhere to be found. Rawls had already argued that justification must come to a stop when we show that a theory explains the judgments of all reasonable persons.[117] He recognized that appeal to agreement was all that one could do:

> one appeals to the voluntary agreement of reasonable men throughout the tradition to mark off the point where one need no longer feel obligated to answer the request for a justification. By carrying our justification this far we have done all that can be done; and the moral skeptic is using the word 'justification' in such a way that it is logically impossible to satisfy him.[118]

Nonetheless, Rawls felt dissatisfied with this state of ethical argument. Appeal to the tradition of ethical thought was the "prima facie evidence that the principles have some moral validity" but "one would hardly wish to be satisfied with the appeal alone."[119]

Linguistic philosophy's notion of absurdity helped Rawls strengthen this argument by adding that the decision to stop the argument on these particular principles, far from being arbitrary, was in some sense natural because the denial of these principles was odd and incomprehensible. Calling this argument "justification by reason," he now claimed that "one ought to show that the principles are reasonable; and that the denial of them either leads to

[114] Rawls, "Remarks on Ethics," 4.3, 4.4., 4.12, 4.15.
[115] Rawls, "Remarks on Ethics," 4.6-4.11.
[116] See, for example, Rawls, "Remarks on Ethics," 5.2.
[117] Rawls, "Remarks on Ethics," 8.25.
[118] Rawls, "Remarks on Ethics," 5.2.
[119] Rawls, "Remarks on Ethics," 5.1.

absurdity or promotes a situation which reasonable men cannot accept."[120] But to be so crucial to our thought that their rejection is absurd, ethical principles had to be very general, as Rawls's principles were.[121]

To exhibit the absurdity of denying his ethical principles, Rawls resorted to an example of a person who "lives nonsense": a person who, rejecting the common sense principles, is also forced to abandon our common sense judgments and ways of living.[122] He took the example from conversations with Malcolm, who, in turn, gleaned it from Samuel Butler's novel *Erewhon*.[123] The idea was to show that some of our beliefs are so crucial that a society which rejected them would be so incomprehensible that it is nowhere to be found (hence the novel's title, which, despite the misplaced "w" and "h," is meant to read "nowhere" in reverse). Rawls used Butler's argument to provide further justification for his first principle, that involuntary actions should not be treated as indicative of the moral worth of our character. Butler depicted a society which acted against this principle: people were put to prison for being sick and sent to the hospital for committing a crime.[124] Other interpersonal relations changed accordingly: for example, sick people were met with moral indignation. As Malcolm had argued in conversation with Rawls at Cornell, we cannot correctly call the judge morally indignant of the defendant's illness because one of our key criteria for moral indignation is voluntariness of action.[125] Adding that this conclusion reflects our stable attitudes, Rawls implied that to reject these attitudes would be to live nonsense.[126]

Although "justification by reason" was undoubtedly a new argument, Rawls portrayed it as part of the logical positivist framework. To meet the demands of the empirical ethical theory, he argued that the appeal to absurdity was not an appeal to a standard. "It is not such an appeal at all," he wrote. Whether something is absurd or not is "a question of fact," and this fact was established by seeing whether the quality of absurdity is agreed upon by "the voluntary agreement of reasonable men."[127] Like the notions of the "reasonable person" and "rational judgments," "absurdity" was defined in such a way that it left the assumptions of an empirical theory very broad and inclusive. In 1947, Rawls's theory was meant to be thoroughly empirical.

Dilemmas of a Positivist

Logical positivism provided Rawls with a seemingly fruitful conception of ethical theory. During the first two years of graduate school he outlined a feasible account of objectivity and made important steps in showing that ethical judgments are indeed objective. Despite this

[120] Rawls, "Remarks on Ethics," 5.1.
[121] In his description of the principles, Rawls refers to Kant, but his actual defense of these principles relies on linguistic philosophy's notion of absurdity. Rawls, "Remarks on Ethics," 5.14-5.15.
[122] Rawls, "Remarks on Ethics," 5.1.
[123] Rawls, "Remarks on Ethics," 8.18. Samuel Butler, *Erewhon, or, Over the Range* (London: Ballantyne Press, 1880).
[124] Butler, *Erewhon*, 71-83.
[125] Rawls, "Remarks on Ethics," 8.18.
[126] Rawls, "Remarks on Ethics," 8.19.
[127] Rawls, "Remarks on Ethics," 5.7.

innovation and fruitfulness, Rawls's approach to ethical inquiry showed signs of future dilemmas already in "Remarks on Ethics." In this section, I will outline three such dilemmas: Rawls's inability to explain the reasons for the agreement of reasonable persons without retreating to moral realism, his overly homogenous and mechanistic account of human judgment, and his attempt to maintain the hypothesis that reasonable persons agree in their ethical judgments in spite of apparent disagreement. These dilemmas would shape many of Rawls's questions which I will discuss in the following chapters.

Rawls's most immediate dilemma was his inability to explain the expected agreement of reasonable persons without retreating to moral realism. In 1947, his account of objectivity and his argument against emotivists lacked the necessary conceptual background. Emotivists had argued that disagreement in ethical judgments is evidence of their subjectivity. Rawls reversed this claim, arguing that agreement in ethical judgments "indicates the objectivity and validity" of these judgments.[128] This argument by itself did not defeat the emotivist thesis, as this latter relied on wider claims about ethical judgments and their relation to human emotions.[129] The wider emotivist conceptual background explained why agreement in ethical judgments could not obtain, or, if it did obtain, why it was a fortuitous and accidental fact. To make his argument stand against this broader emotivist edifice, Rawls needed a wide conceptual framework of his own. He needed to explain why he expected all reasonable persons to agree in their ethical judgments.

Between 1947 and 1951, Rawls reluctantly appealed to logical positivism to detail this framework. Logical positivists explained agreement of normal scientific observers claiming that protocol statements were reports of a physical reality and that normal scientific observers had sufficiently similar epistemological apparatuses. This argument pushed Rawls dangerously close to affirming a physical existence of ethical qualities. His 1947 rejection of Dewitt Parker's claim that the "world is valueless apart from man" suggests that Rawls thought of ethical qualities as not dependent on the human mind.[130] Yet he refused to detail the nature of these values, brushing aside claims that the moral quality discerned by ethical insight existed in the physical object or in the brain.[131] According to Rawls, ethical theory did not need to take sides on this question as it relied on the agreement of reasonable men and the absurdity of denying the proposed principles. As he put it as late as 1952:

> it does not matter at all what the causes of prudential decisions actually are. They may be the result of intuition of non-naturalistic value qualities and so on. All that we attempt here is to explicate these choices; and if the use of the principles would lead to the same choices that competent counselors recommend, then that

[128] Rawls, "Remarks on Ethics," 1.2., 1.1, 5.19. Rawls clearly stated that agreement did not constitute objectivity. See Ibid, 1.2.
[129] Rawls argued against the analytic-synthetic distinction in "Remarks on Ethics," 5.21-5.22, writing that "The grounds for this dictum is nowhere stated. It is one of the crudest opinions ever held in philosophy" and that "It can be shown, I think, that this division [between analytic and synthetic] itself is analytic, and consequently so is their whole theory of ethics. Therefore it is not properly an ethical theory at all." Ibid, 5.22.
[130] Rawls, "Remarks on Ethics," 8.10-8.11.
[131] Rawls, "Remarks on Ethics," 4.18-4.19.

> in itself is sufficient for our purposes, so far as we are concerned with explication.[132]

By 1950, however, Rawls's argument veered closer to moral realism. It is not altogether surprising, as moral realism is a view typically related to the logical positivist conceptions of science. Borrowing this conception of science, Rawls also reluctantly borrowed the related moral realist position. His dissertation shows that, despite the constant claims that questions about the nature of ethical qualities are irrelevant to ethics, Rawls felt the need to posit objectively existing qualities. Now calling them "objective factor" or "objective moral fact," he argued that these qualities determined ethical judgments of reasonable persons, thereby implying that they were mind-independent.[133] The notion stemmed from his earlier analogy between perception and ethical insight:

> just as our common perceptions are caused by, and controlled by, an objective order to events, so we have some reason to think that there is a common objective moral fact which causes and controls our moral judgments….[134]

In contrast to 1947, in 1950 the notions of "reasonable person" and "rational judgment" were defined in relation to this "objective factor." The reasonable person was now more sharply modeled on knowing, and constraints imposed on this definition were interpreted as "necessary conditions for the reasonable expectation that a given person may come to know something" and essential for the "knowledge-getting process" and "finding the truth."[135] The rational judgments were also modeled around the objective factor, which was expected to "evidence itself through the complex of different cultural and personal backgrounds."[136] Analysis of ethical judgments was now not solely the search for agreement among reasonable persons, but also the grasping of this objective factor: "we cannot locate this factor unless we go directly to spontaneous judgments as defined above."[137]

As in 1947, Rawls chose to not detail the nature of the objective factor: "I leave aside the question as to how this common objective moral fact is to be interpreted."[138] Yet given his repeated appeals to this notion, by 1950 it was becoming evident that Rawls felt the need to elaborate this broader conceptual framework in order to explain the agreement of reasonable persons. Relying on logical positivism, he was forced into analogies between ethical insight and perception and led toward the picture of ethical qualities that drew their objectivity from an "objective factor" residing in an ethical situation. In short, modeling ethics on science forced Rawls toward affirming some type of mind-independent existence of ethical qualities and therefore to a form of intuitionism which he had rejected. His way out of this dilemma was to refuse to give any content to the notion, but this way came at the cost of depriving him of the

---

[132] John Rawls, "Theory of Goods" [1952]. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 2. "Considerations of Some Objections," 1i.
[133] Rawls, "Grounds of Ethical Knowledge," 47-48, 277-78.
[134] Rawls, "Grounds of Ethical Knowledge," 278.
[135] Rawls, "Outline of a Decision Procedure in Ethics," in Rawls, *Collected Papers*, 4. Rawls, "Grounds of Ethical Knowledge," 43, 36.
[136] Rawls, "Grounds of Ethical Knowledge," 49.
[137] Rawls, "Grounds of Ethical Knowledge," 48-49.
[138] Rawls, "Grounds of Ethical Knowledge," 278.

broader conceptual background to explain the agreement of reasonable men. These problems would contribute to Rawls's eventual modification of positivism.

Rawls's second dilemma would not emerge until the early 1950s. It would not be a dilemma in the ordinary sense of the word – something over which Rawls would worry, or even something that he would acknowledge as a difficult problem. But it would be a question which would emerge again and again, and a question to which Rawls would give a rather confident answer. I call it the "homogeneity of judgment" dilemma, as the main worry revolved around the account of ethical judgment which Rawls's argument required. This dilemma stemmed from the uneasy combination of two claims: that the correct principles of a theory predict its subject matter, and that the principles can and should replace the currently used intuitive judgments. The key commitment leading to this dilemma was the logical positivist conception of analysis, or its requirement that ethical principles predict the judgments they explain. As Rawls wrote in 1952, "principles are like functions in mathematics: just as functions applied to numbers, say, yield other numbers; so principles applied to circumstances yield a decision."[139] Combined with the "necessary rules of application," the principles had to predict the actual judgments of reasonable persons: "a person who fully understands [both the principles and the rules of application], will be led, by their use, to employ the term expressing the concept on exactly the same occasions, and in exactly the same way, as that term is used in the data with which the explication is concerned."[140] Rawls specified that the principles were to be so precise that they are "mechanically followed," or used by consciously applying the rules without appeal to intuition.[141]

But it was always a question of how these mechanical rules uncovered by philosophical analysis are related to reasonable persons' judgment. Rawls always held that ethical principles are in some way implicit in our judgments, and that the philosopher had to formalize these implicit "rules and principles."[142] Yet the nature of this implicitness needed clarification. Do these rules determine the judgments of reasonable persons? If so, is it because reasonable persons explicitly follow these rules in their judgments? If the judgments of reasonable persons are actually made for other reasons, how are the rules uncovered by analysis related to these other reasons? These are the questions that would surface again and again, until the publication of *A Theory of Justice*.

Rawls's answer to these questions changed. Between 1947 and 1950, the understanding of implicitness was very formalistic and consistent with logical positivism's commitments: Rawls did not require that the principles be used by reasonable persons in their judgments. As he specified in the dissertation, in formulating the principles we are not looking into "what people intend to assert," or "what is before their minds when making an assertion."[143] Inquiry into this intentional meaning, he thought, was "an unnecessary inquiry which it is possible and helpful to

[139] Rawls, "Theory of Goods," Lecture 7, 7i.
[140] Rawls, "Grounds of Ethical Knowledge," 72.
[141] Rawls, "Grounds of Ethical Knowledge," 73. This view sharply differs from contemporary conceptions of ethical principles that trace their origins to Rawls's work. See, for example, Rawls's student O'Neill's *Justice and Virtue*, 77-90, which defends a more open-ended and non-mechanistic understanding of ethical principles.
[142] Rawls, "Grounds of Ethical Knowledge," 75.
[143] Rawls, "Grounds of Ethical Knowledge," 77.

avoid."[144] Indeed, he went so far as to argue that "an explication may be successful even if it can be established with certainty that everyone would reject it as a statement of what they intend to assert."[145] This view implied that reasonable persons did not have to recognize these principles as their own for these principles to be deemed correct. In theory, as long as reasonable men could understand them, the principles could be mathematical formulas employing no intentional notions. Appropriately, referring to Neurath, Carnap, and Popper's works as examples of such explications Rawls called his view a "logical physicalism so far as it may be applied to ethical theory."[146]

Insofar as Rawls claimed that the ethical principles predicted human judgments almost behavioralistically, without taking intentional meaning into account, he made no claims about the actual human judgment and the complexity of ethical life. In the later years, however, he would claim that the principles can play the role of a "replacement schema": they can be used in lieu of our more intuitive ethical judgments.[147] Combined with other assumptions, this view would imply that reasonable persons would acknowledge these principles as implicit in their judgments; some because they deliberately use these principles in their reasoning and others because they would admit after reflection that they do indeed use these principles. As Rawls would formulate the principles of justice as a "replacement schema," he would encounter problems using the mechanistic model of principles in a world of human affairs. In particular, he would be led to contestable conclusions about the complexity of political life and human judgment. To serve their function in an empirical ethical theory, ethical principles had to predict their subject matter: they had to determine a particular ethical judgment. To do so, there had to be but one ethical principle, or, if there were more, they had to be ordered in terms of importance in case they should conflict with each other. That in turn entailed a contestable conclusion that all reasonable persons make all of their ethical judgments for the same one or more reasons.[148] Thus, adopting the logical positivist conception of axioms and principles, Rawls would be pushed to a view that ethical or political life is simple and tidy in the sense that it can be decided by using a limited number of ordered reasons. We will come back to this dilemma in Chapters Four and Five.

Rawls's final dilemma stemmed from the demands of scientific ethical theory: he had to reconcile his claim that the judgments of all reasonable persons overlap with the fact of apparent disagreement. In the years 1946-50, this uneasy relationship between Rawls's goals and facts of

---

[144] Rawls, "Grounds of Ethical Knowledge," 76f.
[145] Rawls, "Grounds of Ethical Knowledge," 77.
[146] Rawls, "Grounds of Ethical Knowledge," 78. Carl Hempel, "A Purely Syntactical Definition of Confirmation," *Journal of Symbolic Logic* 8 (1943): 122-43; Carl Hempel, "Studies in the Logic of Confirmation," *Mind* 54 (1945): 1-26; Rudolf Carnap, "On Inductive Logic," *Philosophy of Science* 12 (1945): 72-97; Rudolf Carnap, "The Two Concepts of Probability," *Philosophy and Phenomenological Research* 5 (1945): 513-532. The reference to these works is found in Rawls, "Grounds of Ethical Knowledge," 73f.
[147] Rawls, *A Theory of Justice*, 111.
[148] The view that analysis should result in a definition applicable to all cases in which the word is used was widespread among logical positivists and analytic philosophers more broadly. See Moritz Schlick, *Problem of Ethics*, trans. David Rynin (New York: Prentice Hall, 1939), 12; Moore, *Principia Ethica*, ix-x; C.D. Broad, *Five Types of Ethical Theory* (New York: Harcourt, Brace & Co., 1930), 82; W.D. Ross, *The Right and the Good* (Oxford: Clarendon Press, 1930), 90. Rawls would encounter problems with this assumption only in the early 1950s, when he would engage linguistic philosophers such as Stuart Hampshire and Stephen Toulmin. It is also worthwhile to point out that Rawls's students would reject the claim that understanding justice in terms of principles implies the view that all questions of justice are solved by the same set of reasons. See O'Neill, *Justice and Virtue*, 73-90.

life was not a live dilemma for Rawls: interested in uncovering the function of ethical claims, he saw the discrepancy of moral visions as a natural occurrence, explained by contextual factors such as the presence of warlike neighbors. In the later years, when linguistic philosophers would convince him to turn to the intentional meaning of ethical utterances, the dilemma would become pressing for Rawls and he would seek ways to solve it. His main strategy would be to claim that reasonable persons do not have to agree on everything – only on enough to also share a conception of justice. But he would also claim that reasonable persons do not have to agree straightaway – only "after reflection." I will look into these strategies in the remaining chapters.

Conclusion

Rawls's early thought formed in the logical positivist tradition. Modeling ethical inquiry on the logical positivist conception of scientific theory, he presented a "physicalist" ethical theory. The theory was deductive, consisting of axioms or principles that aim to predict its subject matter: ethical judgments. It was also a non-foundational theory: ethical judgments were thought to be objective insofar as all reasonable persons agreed on them. Despite its clear aspirations to be an empirical theory, it was also normative, presenting principles for an internal critique in the universal community of reasonable persons.

This conception of ethical inquiry set Rawls's main tasks between 1946 and 1951: he aimed to show that all ethical judgments do indeed exhibit the stability expected of events that occur for a reason. To that end, he developed two notions: the "reasonable person" and "rational judgment." The results of his two distinct theories in 1946 and 1947 were limited: although both universalistic, the first explained the uses of ethical words whereas the second uncovered background presuppositions of all ethical theories. Between 1946 and 1950, Rawls did not enter difficult ethical debates and suggested no "line of conduct": he thought that difficult ethical questions were unresolvable. These limited aims put the task of *A Theory of Justice* in perspective: aiming to uncover the criteria for practices that are just, Rawls's classic work would far surpass its logical positivist predecessors in its practical aspirations.

The positivist conception of theory also created Rawls's dilemmas and future problems. In the following two decades, he would be forced to defend and modify the homogeneity thesis and clarify the way in which ethical theory was an empirical theory. Most immediately, he was forced to give reasons for expecting the agreement of all reasonable persons, and do so without resorting to the realist notion of the "objective factor" that governed their decisions. Rawls's way out of this problem started in 1947, when he adopted the first concepts from linguistic philosophy. Between 1947 and 1951, this tradition brought about few, if noticeable, changes in his thought. However, the notion of "absurdity" relied on a wider conceptual framework that was more amenable than logical positivism to explaining the agreement of reasonable persons without resorting to moral realism. Its notions of human practice and the form of life would captivate Rawls and, over the next decade, he would shed many of his positivist commitments.

The following two chapters narrate this transformation. They show Rawls drawing on linguistic philosophy to elaborate the conceptual framework that explains the expected convergence of reasonable persons' opinions. Chapter Four narrates Rawls's drawing on

linguistic philosophy's key idea of human practice to present ethical reasoning as a practice governed by shared rules, while Chapter Five outlines Rawls's attempt to explain human agreement by presenting morality as an extension of natural feelings.

4

Ethical Reasoning as a Practice: Themes from Linguistic Philosophy

Rawls's conception of philosophy underwent significant changes in the early 1950s, as he departed for a year at Christ Church, Oxford, where he was influenced by linguistic philosophers and especially by Ludwig Wittgenstein and his students. The folk narrative that sees Rawls as a lone figure reviving political theory hides many interesting developments in mid-20th century political philosophy. One such development was the rise of linguistic philosophy, which set itself against positivism and particularly against its analytic-synthetic distinction.[1] Linguistic philosophers differed among themselves, and, as scholars have argued, we should acknowledge the diversity of Wittgensteinians and ordinary language philosophers.[2] To Rawls, who pondered on these novel arguments from the point of view of his positivist framework, the most convincing current was Wittgensteinian as it was interpreted by Wittgenstein's student Stephen Toulmin. It is not a coincidence that Rawls should favor this current: focusing on the notion of practice or human activity governed by socially accepted rules, Toulmin's Wittgensteinianism overlapped with Rawls's positivism in its search for regularities in normative judgments. Understanding language and reasoning as practices, Wittgensteinian philosophers such as Toulmin saw these rules as constitutive of what counted as an appropriate use of a word, and took explication of these rules as the main task of philosophy.

Early linguistic philosophers, including Ludwig Wittgenstein himself, were not concerned with ethics or political philosophy, but by the early 1950s their students had extended the novel approaches to these disciplines. Seeing ethical reasoning as a practice governed by rules, they sought to spell out its logic and specify what reasons count as good reasons in ethical arguments.[3] Hence the name by which it is known, the "good reasons" approach.[4] This study of ethical reasoning was known for its express anti-foundationalism: linguistic philosophers refused

[1] On the origins of linguistic philosophy and the nature of the linguistic turn, see: Michael Dummett, "Oxford Philosophy" in Michael Dummett, *Truth and Other Enigmas* (Cambridge, MA: Harvard University Press), 431-436; Isaiah Berlin, "Austin and the Early Beginnings of Oxford Philosophy" in *Essays on J.L. Austin* (Oxford: Clarendon Press, 1973), 1-16; P.M.S. Hacker, *Wittgenstein's Place in Twentieth Century Analytic Philosophy* (Oxford: Blackwell Publishers, 1996); Bryan Magee interview with Bernard Williams in "The Spell of Linguistic Philosophy" (Princeton, NJ: Films for the Humanities & Sciences, 2003); Rorty, "Introduction" to *Linguistic Turn*, 1-39. See also Hans Sluga, "What Has History Got to Do with Me? Wittgenstein and Analytic Philosophy," *Inquiry* 41 (1998): 99-121.

[2] Dummett, "Oxford Philosophy."

[3] William Frankena, "Main Trends in Recent Philosophy: Moral Philosophy at Mid-Century," *The Philosophical Review* 60 (1951): 44-55.

[4] For key arguments in the "good reasons" approach, see Toulmin, *Reason in Ethics*; Kurt Baier, "Good Reasons," *Philosophical Studies* 4 (1953): 1-15; Kurt Baier, "Proving a Moral Judgment," *Philosophical Studies* 4 (1953): 33-44; Kai Nielsen, "The 'Good Reasons Approach' and 'Ontological Justifications' of Morality," *The Philosophical Quarterly* 9 (1959): 116-130.

to provide any other justification for ethical reasons than human agreement and regarded such requests for further foundations as impossible to satisfy.

Although Rawls had encountered Wittgenstein's ideas through his teacher Norman Malcolm in the 1940s, they had little impact on his thought at the time. That changed in the early 1950s, as he read works of linguistic philosophers, especially Stephen Toulmin, Stuart Hampshire, and Herbert Hart. Already understanding ethics as a study of ethical judgments, Rawls found many themes from their philosophy convincing, and, indeed, soon adopted their central views, regarding ethical reasoning as a practice and ethical theory as an attempt to elaborate the rules of this practice. This appropriation led to important changes in his thought. First and most notably, he changed his conception of justification: he started insisting that, to justify reasons given in political arguments, one only needed to show that all reasonable persons shared them. This helped sever any links with moral realism to which positivism seemed to lead Rawls. Second, the notion of practice influenced Rawls's political theory: he restricted the subject matter of justice to the major institutions or practices of human society.

Equally important are ways in which Rawls modified his grandiose positivist theory all to respond to the objections of linguistic philosophers. Linguistic philosophy threatened the positivist conception of theory. If it suggested that all situations of justice are decided for the same reasons, linguistic philosophers appealed to the slogan "meaning is use" to argue that different cases are decided for different reasons. And if positivism offered a mechanistic conception of judgment whereby philosophy presents reasons (principles) which are then used to decide cases in practice, linguistic philosophers thought that philosophy could only produce paradigmatic examples of good judgments. Rawls saw these differences between the two approaches as significant, but he came up with his own answers to these potential problems. Thus, while he endorsed the slogan "meaning is use," he did not practice linguistic philosophy's contextualism and instead aimed to explain our political judgments using the methods of decision theory. Similarly, he still aimed to build a theory of ethical judgments – to provide a list of reasons relevant for deciding questions of justice – but now acknowledged that principles would guide practice only in a "loose" manner. These modifications are significant; while Rawls kept the description of the positivist theory intact, he significantly changed his mind about the extent to which these goals can actually be achieved.

In this chapter, I want to narrate the changes in Rawls's positivism that took place between 1950 and 1954, and show the importance of linguistic philosophy in driving these changes. I divide the narrative into four parts. In the first, I outline the main themes of early linguistic philosophy at Cambridge and Oxford: language as a practice, meaning as use, family resemblance, and the notion of 'limiting questions' or the 'stopping point of justification.' In the second, I outline the character of linguistic philosophy in the early 1950s, as Rawls encountered it in the writings of Toulmin, Hampshire and Hart. In the third part, emphasizing Rawls's changing conceptions of justification and ethical principles, as well as his delineation of the subject of justice, I narrate the transformation of his positivism between 1950 and 1952. In the final part that covers Rawls's stay at Oxford and early years at Cornell (1952-54), I describe the initial versions of the hypothetical reasoning game that, in *A Theory of Justice*, would become known as the "original position."

Themes from the Early Linguistic Philosophy

Linguistic philosophy arose against the logical positivist tradition with its analytic-synthetic distinction and its reductionist attempt to show that all meaningful statements collapse into statements that are either analytic or synthetic. According to this picture, to be convincing, ethical reasons also had to be reducible to either synthetic or analytic statements. Linguistic philosophy rejected this reductionism and insisted that, if one attends to our actual appeal to these reasons, one sees that not only do we not ask to justify them further, but also that such requests cannot be meaningfully made. To defend this view, linguistic philosophers introduced the notions of a 'practice' or 'language game,' as well as a 'function' that such practices have in the larger range of human activities. Relying on these notions, it insisted that the criteria for good and bad reasons are given by our shared understandings and argued that no further reasons for these shared understandings can be given. Refusing to ground reasoning – whether in ethics or any other intellectual inquiry – in epistemologically secure foundations, linguistic philosophy was thus not only anti-foundationalist but also anti-reductionist.

Ludwig Wittgenstein's later thought was by far the most important source of this new tradition in philosophy. It centered on the view of language as a practice, or a set of activities best understood by two features: they were governed by shared understandings or what he called "conventions," and, playing a role in human activities, they were used to do certain things, achieve certain purposes.[5] First, then, shared understandings constituted what counted as the right use of a word: as certain moves in games count as "false moves," or moves that violate the rules of the game, so certain uses of words count as improper uses. [6] Whether a particular use of a word is appropriate depends on shared understandings or conventions. These conventions, he thought, are implicit, often imprecise, but, for them to be conventions, they have to be shared.[7] Even though he allowed that human beings can disagree about particular statements, Wittgenstein expected human beings to share these conventions, or "agree in the *language* they use."[8] Second, practices played a role in human activities: various words were used for different purposes, to do different things. To emphasize this role of practices, Wittgenstein coined the term "language-game": "the term 'language-*game*' is meant to bring into prominence the fact that the *speaking* of language is part of an activity, or of a form of life."[9] He insisted that "there are different *kinds of word*" – words that have different functions – and his examples of practices reflected these different uses of words: giving orders, reporting events, forming and testing a hypothesis, play-acting, and making a joke, were all considered practices.[10] Other practices were broader and included academic disciplines, such as chemistry and calculus. [11] Nonetheless, the underlying idea was the same: within any practice, some moves can, and some cannot be made.

Wittgenstein insisted that the reductionism advocated by positivism was wrong in two important respects. Most importantly, reductionism such as Russell's attempt to reduce all

---

[5] Wittgenstein, *Philosophical Investigations*, §§355, 7.
[6] Wittgenstein, *Philosophical Investigations*, §§355, 345.
[7] Hence Wittgenstein's argument against the possibility of a private language, where criteria for correct application of a word are known to the user of that language alone. See Wittgenstein, *Philosophical Investigations*, §§269-75.
[8] Wittgenstein, *Philosophical Investigations*, §241.
[9] Wittgenstein, *Philosophical Investigations*, §23.
[10] Wittgenstein, *Philosophical Investigations*, §§7, 17. Emphasis original.
[11] Wittgenstein, *Philosophical Investigations*, §§18, 23.

sciences to logic assumed that, to be credible, reasons generated within a language-game needed support in some more basic language-game. This, Wittgenstein thought, was a mistake. His argument rested on an unsaid distinction between reasons given from within a language-game and reasons given in support of that language-game.[12] Bringing up an example of a schoolboy who questioned the existence of all things, Wittgenstein wrote that "this pupil has not learned how to ask questions. He has not learned the game that we are trying to teach him."[13] In this game, the existence of things is one of the background assumptions that is taken for granted, and the pupil's doubt "isn't one of the doubts in our game."[14] Wittgenstein did not think that we could give no reasons to convince the schoolboy about the appropriateness of our game – its success in explaining our world would be one of them – but he did think that the boy's question, if understood literally, could not be asked.[15]

The notions of 'practice' and questions that cannot be asked were the core parts of Wittgenstein's understanding of philosophy. According to him, philosophical problems stem from conceptual confusions, which consist in extending a word from its original language game to a language game with different conceptual connections. Each proposition or question, he argued, has its own "grammar," or a set of propositions to which it relates.[16] Removing the concept from its original home leads to a conceptual confusion, or "nonsense" – statements that have no meaning, that cannot be connected to other concepts in any recognizable way.[17] To solve philosophical problems, Wittgenstein thought, a philosopher had to dispel conceptual confusion by reminding the interlocutor how a word in question is used in its original language game: "When philosophers use a word – 'knowledge,' 'being,' 'object,' 'I,' 'proposition,' 'name' – and try to grasp the *essence* of the thing, one must always ask oneself: is the word ever actually used in this way in the language-game which is its original home? – What *we* do is to bring words back from their metaphysical to their everyday use."[18] Importantly, Wittgenstein thought that we already have everything we need in order to solve philosophical puzzles, and that showing the word's original language game would make it clear that its relevant extension is illegitimate: "The problems are solved, not by giving new information, but by arranging what we have always known."[19]

Wittgenstein's arguments in epistemology are a good example of rejecting requests for further justification. Assessing the positivists' insistence that all our knowledge be inferred from sense experience, he claimed that, if we attend to the uses of relevant concepts 'inference' and 'evidence,' we will see that we infer only in very specific contexts, such as when the object in question is not in sight and our only evidence for its existence are the shadows that it purportedly

[12] For a recent version of this argument, from which I also borrow, see Peter Lamarque, "Wittgenstein, Literature, and the Idea of a Practice," *British Journal of Aesthetics* 50 (2010): 375-88.
[13] Ludwig Wittgenstein, *On Certainty*, edited by G.E.M. Anscombe and G.H. von Wright and translated by Denis Paul and G.E.M. Anscombe (New York: Harper Torchbooks, 1969), §315.
[14] Wittgenstein, *On Certainty*, §317.
[15] Wittgenstein, *Philosophical Investigations*, §§320, 324.
[16] Wittgenstein, *Philosophical Investigations*, §353.
[17] For an example of a proposition the grammar of which is unclear, see Wittgenstein, *Philosophical Investigations*, Part II, 221.
[18] Wittgenstein, *Philosophical Investigations*, §116.
[19] Wittgenstein, *Philosophical Investigations*, §109.

casts.[20] When the object in question is in sight, no inference needs to be made. Indeed, a request for evidence in this case would be paradoxical, and that is because the grammar of the concept 'evidence' is violated. 'Giving evidence' or 'justifying' was a language game of its own – a language game that could be played only in certain circumstances. To always ask for justification is to misuse the word. "To use a word without a justification," he wrote calling attention to this point, "does not mean to use it without right."[21] This idea that one needs to give reasons or justifications only in certain circumstances would be particularly influential on the ethical writings of later linguistic philosophers.

Reductionism was, according to Wittgenstein, mistaken in a second way as well: it led us to expect that beneath the multiplicity of language games in which words like 'justice' are used, there is some common language, common meaning to all uses of 'justice.' Yet, attentiveness to these various language games would, he argued, reveal that its logical grammar differs across contexts. Taking games as an example, he argued that there was no one feature common to all of them: "For if you look at them you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that."[22] These "criss-crossing" and "overlapping" similarities reminded Wittgenstein of family members who all had many traits in common but did not necessarily share any one trait; hence the name, "family resemblances."[23] This last theme of 'family resemblances' would become particularly important for ethics, and, later, John Rawls.

These characteristic themes of linguistic philosophy – its understanding of reasoning as a practice, its distinction between reasons generated by the practice and reasons in support of that practice, its refusal to give reasons for these latter, its notion of "nonsense" or questions that cannot be asked, its attention to the uses of words – soon became widespread at Oxford and Cambridge. Part of this influence was undoubtedly due to Wittgenstein: his students, his illicitly reproduced lecture notes, known as the *Blue Book* [1933-34] and the *Brown Book* [1934-35], and meetings with him made for many conversions to linguistic philosophy.[24] Perhaps the most famous example is Gilbert Ryle, who, upon meeting Wittgenstein in 1929, abandoned his initially Russellian conception of philosophy for recognizably Wittgensteinian commitments. He now saw the source of philosophical mistakes in an illegitimate extension of the concept from its original to foreign contexts – a move he called a "category mistake."[25] He argued that, since the grammar of different concepts – its presuppositions and entailments – is also different, this mistake gives rise to philosophical puzzles that can be disentangled by examining the actual reasoning and charting the "logical powers of ideas."[26]

---

[20] Wittgenstein, *Philosophical Investigations*, §§472-80.
[21] Wittgenstein, *Philosophical Investigations*, §289.
[22] Wittgenstein, *Philosophical Investigations*, §66. Emphasis original, as throughout the dissertation.
[23] Wittgenstein, *Philosophical Investigations*, §§66, 67.
[24] For historical narratives, see Lynd Forguson, "Oxford and the 'Epidemic' of Ordinary Language Philosophy" *The Monist* 84 (2001), 332; Berlin, "Austin and Oxford Philosophy," 11; Hacker, *Wittgenstein's Place in Analytic Philosophy*, 86. For examples of work by Wittgenstein's students, see Alice Ambrose, "Finitism in Mathematics I," *Mind* 44 (1935), 186-203; Alice Ambrose, "Finitism in Mathematics II," *Mind* 44 (1935), 317-340; John Wisdom, "Philosophical Perplexity," *Proceedings of the Aristotelian Society* 37 (1936), 71-88.
[25] Gilbert Ryle, "Philosophical Arguments," in Gilbert Ryle, *Collected Papers*, vol. II (London: Hutchison & Co, 1971), 200. See also Scott Soames, *Philosophical Analysis in the Twentieth Century*, vol. II (Princeton: Princeton University Press, 2003), 94.
[26] Ryle, "Philosophical Arguments," 201.

Although not all of the linguistic tradition started with Wittgenstein, and certainly not all of its members agreed on all issues, even philosophers on the margins of this movement shared many of its themes. John Austin is one such example: although he lacked the notion of 'practice,' his thought stemmed from themes characteristic of linguistic philosophy: the notion of nonsensical questions, or questions that cannot be asked, attention to the use of words, and reliance on shared understandings or conventions. Thus, he thought that one source of philosophical mistakes was departures from any context whatever, or the mistake of "asking nothing in particular."[27] This tendency to generalize – typical of philosophers but not of ordinary persons – leads us to ask general questions like "what is the meaning of a word?." Yet, since departing from any context, such questions ask "nothing in particular," they are "pseudo," or nonsensical questions. [28] Nonsensical questions also included those that extend a concept "to cases that have by now too tenuous a relation to the model case."[29] Austin's later writings centered on this latter source of mistake; *Sense and Sensibilia*, in which, much like Wittgenstein, he argued that Ayer's request for evidence is an illegitimate extension of the concept 'evidence,' is an excellent example of this type of reasoning.[30]

Like Wittgenstein, to show that certain questions or positions are nonsensical, Austin relied on authoritative appeals to our knowledge of the ordinary usage of relevant expressions, or "*what we should say when*, and so why and what we should mean by it."[31] Thus, in his arguments against the generalizing philosopher, Austin often appealed to the plain man and his puzzlement at the philosopher's questions and answers.[32] Despite an explicit denial that ordinary language serves as the "last word," he defended reliance on our linguistic sense.[33] First, he thought, we all simply agreed on this usage: "the more we imagine the situation in detail, with a background story," he wrote, "the less we find we disagree about what we should say."[34] Second, more controversially, he argued that ordinary language acts as a depository of moral knowledge: it "embodies all the distinctions men have found worth drawing, and the connections they have found worth marking, in the lifetimes of many generations."[35] Although this last argument starkly distinguished Austin from Wittgenstein, his work exhibited other themes of linguistic philosophy: its emphasis on the use of words, its characterization of these problems as violations of the proper conceptual grammar, and, finally, its reliance on this ordinary usage to draw limits beyond which questions could not be meaningfully raised.

---

[27] John .L. Austin, "The Meaning of a Word" in J.L. Austin, *Philosophical Papers* (Oxford: Clarendon Press, 1961), 26-27.
[28] Austin, *Philosophical Papers*, 25-27.
[29] John L. Austin, "A Plea for Excuses" in Austin, *Philosophical Papers*, 150-51.
[30] John L. Austin, *Sense and Sensibilia*, ed. G.J. Warnock (Oxford: Oxford University Press, 1962).
[31] Austin, "Plea for Excuses," 129. Emphasis original.
[32] See, for example, Austin "Meaning of a Word," 27.
[33] Austin denied that ordinary language is the final court of appeal "partly because ordinary language is not applicable to extraordinary situations, partly because it is sometimes loose and sometimes not uniform, and, finally, because it contains not only insights but also superstition, error, and "fantasy of all kinds." Austin, "Plea for Excuses," 131-33, 137.
[34] Austin, "Plea for Excuses," 132.
[35] Austin, "Plea for Excuses," 132, 130.

Themes from Linguistic Philosophy in the early 1950s

By the early 1950s, the students of Wittgenstein, Ryle and Austin started bringing linguistic philosophy's themes to bear on questions in ethics and politics. Stephen Toulmin, Wittgenstein's student at Cambridge in the late 1940s, worked on ethics and the philosophy of science, while Oxford thinkers Isaiah Berlin, Stuart Hampshire, and Herbert Hart were publishing on ethics, politics, and history of political thought. This younger generation of linguistic philosophers would continue the core themes of the tradition and, in particular, insist on its opposition to theory construction in ethics and politics.

Most notably, these later linguistic philosophers were expressly non-reductionist. Viewing ethical reasoning as a practice, they argued that this practice governs what counts as a good reason in ethical arguments. Reasons, they insisted, are always given within a practice, but the practice of ethical reasoning itself cannot be justified further. This argument was most forcefully expressed in Toulmin's dissertation, filed in 1948 and published in 1950 as *An Examination of the Place of Reason in Ethics*. It was the first book-length treatment of ethics from a recognizably linguistic perspective, but a perspective that was also distinct from his teachers'. Although Toulmin borrowed Wittgenstein's understanding of language as a human practice, his innovation was to argue that ethical reasoning is also a practice and that, like other practices, it has its own rules, and, therefore, criteria for distinguishing between good and bad reasons: "good reasons and bad reasons, correct and incorrect inferences, sound and unsound arguments, all are decided in this case by the rule of the game."[36] This logic, he wrote relying on another Wittgensteinian theme, was determined by the larger function of ethical reasoning in human activities, which, Toulmin thought, was "correlat[ing] our feelings and behavior in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible."[37] Ethical reasons were therefore classified as good or bad by how well they managed to adjudicate between everyone's conflicting aims and desires.

Like the preceding linguistic philosophers, Toulmin denied that one could justify ethical reasoning in any stronger way than showing that it actually fulfils its function and adjudicates conflicts between incompatible desires. He distinguished between justifying within a practice, and justifying the practice itself. Thus, he thought, ethical reasoning, having a logic of its own, provided a way to justify actions and, to a limited extent, practices of a society. One could justify individual actions by appealing to reasons set by a practice: an institution of promising, for example, gives us a reason – though not necessarily a sufficient reason – to do what we promised.[38] One could also justify practices that fall under the purview of ethical reasoning, although such justification, Toulmin thought, was much more limited: one could argue for one practice over another if the proposed new practice could be shown to eliminate the problems of the previous practice without changing anything else, or if it could be shown that a new practice would be more congenial to other practices of our society.[39]

[36] Toulmin, *Reason in Ethics*, 81.
[37] Toulmin, *Reason in Ethics*, 137.
[38] Toulmin, *Reason in Ethics*, 148.
[39] Toulmin, *Reason in Ethics*, 148-52.

However, Toulmin emphasized that ethical reasoning itself could not be given further foundations. Once we have considered all the good reasons for fulfilling our promise and have decided that doing so is the right thing to do, he argued, we have all the reasons to fulfill our promise. The further question, 'But why should I do what I promised?' or, more broadly, 'But why should I do what is right?' can no longer be asked because ethical reasoning cannot provide any further reasons.[40] One could give reasons that go beyond ethics, such as expediency or authority, but these reasons, Toulmin thought, were not appropriate for ethics.[41] In a Wittgensteinian manner, he argued that requests for further foundations extend ethical reasoning beyond its original home: "as a consequence of the ways in which we employ the words concerned, and of the purpose which [ethical] questions … serve, there is logically no place in such a situation for this question – taken literally."[42] Since these unanswerable questions revealed the limits of ethical reasoning, Toulmin called them "limiting questions."[43] The best way to deal with them, he thought, was to address the motives and doubts from which these confused questions arise and to try to dispel the question by explaining to the questioner that the origins of the notions 'right' and 'obligation' are "such as to make the sentence 'One ought to do what is right', a truism."[44] Thus, Toulmin regarded it an impossible task to give a "*general* answer to the question, 'What makes some ethical reasoning 'good' and some ethical arguments 'valid'?"; like other linguistic philosophers, he refused to provide further foundations for all reasons in ethics.[45]

The express anti-foundationalism and anti-reductionism became a common commitment of many philosophers at Cambridge and Oxford. Toulmin's book played an important part in this trend: naming his approach the "toulminian conception of the logic of justification," some of his followers similarly argued that ethics could only provide reasons that are generally accepted as good, that it could not give any further foundations for these reasons, and that such requests for further reasons were confusions of "logical cupboards."[46] Others arrived at anti-foundationalism independently. Ryle's student Kurt Baier, for example, argued that one could provide paradigmatic examples of good reasons in ethics, but that, apart from showing how "irrational" it would be to reject these reasons, philosophers could not give any stronger argument in ethics.[47] Stuart Hampshire's approach to philosophy was similarly characteristic of linguistic philosophy. At its core was a conception of 'logical nonsense,' encompassing expressions that, "though linguistically and grammatically correct, [contravene] some implicit rule of use of the terms involved."[48] Faced with examples of nonsense, philosophers could dispel them by showing that they were logically meaningless.[49] That was all they could do without creating further nonsense: "I have tried to show that philosophical explanation could not 'take one any further,' because 'any further' would be nonsense; the complaint of the imaginary metaphysical reader would

[40] Toulmin, *Reason in Ethics*, 202-3.
[41] Toulmin, *Reason in Ethics*, 163.
[42] Toulmin, *Reason in Ethics*, 203.
[43] Toulmin, *Reason in Ethics*, 204-212.
[44] Toulmin, *Reason in Ethics*, 208.
[45] Toulmin, *Reason in Ethics*, 161.
[46] Nielsen, "The 'Good Reasons Approach'," : 117, 121.
[47] Baier, "Good Reasons"; Baier, "Proving a Moral Judgment."
[48] Stuart Hampshire, "Logical Necessity," *Philosophy* 23 (1948): 339.
[49] Hampshire, "Logical Necessity," 343.

therefore be unreasonable."[50] In ethics, he thought, one could similarly clarify the use of ethical terms, but "no argument can show that B *must* use the criteria which A uses and so must attach the same meaning (in this sense) to moral terms as A."[51] Hampshire was one of the linguistic philosophers who allowed for disagreement in ethical arguments even after all reasons have been given; in such cases, he argued, one has to make a decision: "Between two consistently applied terminologies, whether in theoretical science or in moral decision, ultimately we must simply choose; we can give reasons for our choice, but not reasons for reasons for … *ad infinitum*."[52]

In addition to its anti-foundationalism, linguistic philosophy in the 1950s adamantly rejected constructions of theories in ethics and politics. In its stead, they put forth a philosophy of example. This opposition to theories stemmed from its view of ethical judgment: resting its case on examples, it argued that we make decisions about justice (for instance) appealing to different reasons. Thus Hampshire wrote that concepts such as 'justice' could not be defined in terms of one or several reasons, since such formulations wrongly assume that there must be a "single sufficient reason from which I always and necessarily derive my judgment."[53] This argument was echoed by Herbert Hart, eventually Oxford Chair of Jurisprudence and one among the participants in the Saturday morning discussions with Austin. Hart also concentrated on the ordinary uses of words and, relying on this usage, argued that statements of necessary and sufficient conditions for the correct application of words were usually flawed. Concepts, Hart argued, were "defeasible," as "any set of conditions [for the correct use of a word] may be adequate in some cases but not in others."[54] This conception of ethical reasoning had radical implications for philosophy's usefulness to the practice of politics: both Hampshire and Hart argued that a philosopher burdened with the task of explaining "justice" could only refer to the "leading cases on the subject, coupled with the use of the word 'etcetera,'" and that these paradigmatic examples would be useful in practice by "direct[ing] attention to further known facts as relevant to a judgment."[55] This conception of ethics and its relation to practice stood in sharp contrast to Rawls's positivist structure, which aimed to explain the concept of justice precisely in terms of several principles, and which thought to contribute to practical reasoning not by offering examples of good judgment, but by providing principles from which one could deduce particular practical judgments.

In sum, starting from their conception of ethical reasoning as a human practice, linguistic philosophers arrived at their anti-foundationalism or refusal to entertain requests for further, more stable foundations for reasons in ethical arguments. They argued that such requests stem from a misunderstanding of ethical reasoning, and that at best one could dispel the motivations from which such confused requests arise. Except for Hampshire, these anti-foundationalist philosophers expected that the demonstration of nonsense as nonsense would be convincing to all, including those who produced it; as Wittgenstein wrote, "if one tried to advance *theses* in philosophy, it would never be possible to debate them, because everyone would agree to

[50] Hampshire, "Logical Necessity," 340.
[51] Stuart Hampshire, "Fallacies in Moral Philosophy," *Mind* 58 (1949): 478.
[52] Hampshire, "Fallacies in Moral Philosophy," 478.
[53] Hampshire, "Fallacies in Moral Philosophy," 481.
[54] H.L.A. Hart, "The Ascription of Responsibility and Rights," *Proceedings of the Aristotelian Society* 49 (1948-49): 174, 181.
[55] Hart, "Ascription of Responsibility and Rights," 173-74; Hampshire, "Fallacies in Moral Philosophy," 481.

them."[56] Some elucidation – perhaps arguments by analogy, such as that to a game – was needed, but, in principle, convergence around the logical grammar of concepts was taken for granted. When it was questioned, as in the case of Austin, the breadth of agreement was circumscribed, but a new standard was not produced. As James Opie Urmson, an Oxford philosopher about to visit Princeton as professor in 1950-51, remarked, if agreement around the criteria for good horses, say, fails to obtain, "I do not know what one can do about it. All co-operative activities, all uses of language, must start from some agreed point."[57] Similarly, linguistic philosophers thought that attention to the uses of ethical language and instances of ethical judgment would show that different situations of justice require different reasons, and that they cannot all be summarized in terms of several reasons which are always present and always sufficient to determine our judgment. In that regard linguistic philosophers were pluralists, but they expected that everyone would still agree on this variety of reasons and their weights in different situations. Many themes characteristic of this new conception of philosophy would help or trouble Rawls, but its anti-foundationalism and its rejection of theory-construction would be particularly influential.

Rawls's Early Encounters with Linguistic Philosophy

Linguistic philosophy had many connecting points with Rawls's early positivist thought: throughout the late 1940s, he took actual political judgments as philosophical data and attempted to uncover reasons for which those judgments are made. In Rawls's own broad sense of the word, both positivism and linguistic philosophy saw ethics as an "empirical inquiry."[58] These connecting points provided him with a smooth transition to many linguistic philosophy's themes, and, by the end of 1954, Rawls had already appropriated its center-piece view of ethical reasoning as a practice, its belief that this reasoning has its own logic, that the task of philosophy is to make that logic explicit, and its conception of justification that relies on human agreement. These commitments would lead Rawls in fruitful directions, such as restricting the subject matter of justice to the main social practices of a society, but they would also require important changes in his conceptions of principles.

The first motives of linguistic philosophy in Rawls's thought appeared together with references to Toulmin's *The Place of Reason in Ethics*, Hampshire's "Fallacies in Moral Philosophy" [1948], and William Frankena's overview article "Main Trends in Recent Philosophy" [1951]. Undoubtedly, Urmson's visiting professorship at Princeton in the academic year 1950-51 also played an important part. While at Princeton, Urmson lectured on ethics, "with a view to determining the nature of ethical problems and the criteria for their adequate solution."[59] A year later, when Rawls taught the same course to Princeton undergraduates, his views were already significantly modified in response to the mentioned themes of linguistic philosophy. Between 1950 and 1952, three important developments took place in Rawls's

[56] Wittgenstein, *Philosophical Investigations*, §128.
[57] J. O. Urmson, "On Grading," *Mind* 59 (1950): 169.
[58] Rawls, "Nature of Ethical Theory," 7-8; John Rawls, "On Explication. Oxford" (1952-53). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 18, 77.
[59] Princeton University Course Catalog, 1950-51 (Princeton: Princeton University Press, 1950), 259.

thought: he elaborated a new conception of justification, modified his account of ethical principles, and employed the concept of a practice to restrict the subject of justice to the basic institutions of our society.

Three sets of writings are helpful in exploring the changes that take place between 1950 and 1952: Rawls's 1950 review of Toulmin's *The Place of Reason in Ethics*, "On Values," a paper written most likely in 1951, and the mentioned 1952 Princeton lectures that consist of two files, "Ethics and Its Reasoning" and "Diseases of Ethical Reasoning."[60] By the time Rawls left for Oxford in 1952, the most important changes in his thought had already taken place, and the year-long stay, together with the first years at Cornell, by which Rawls was hired in 1953, brought developments but no sweeping changes. For this reason, if the reformulation of the same idea does not hide an important conceptual advance, I will sometimes use Rawls's later writings – 1953 notes "On Explication" and "Oxford Notes: Spring 1953" (hereafter "Oxford Notes") – to explain the earliest changes in Rawls's thought.[61]

Rawls's new way of looking at justification was centered on viewing ethical reasoning as a practice. As he wrote in "Ethics and Its Reasoning," "reasoning is an activity. It is something that men do."[62] To bring to light the implications of this comparison, he introduced the analogy to a game: while reasoning is not a game, he wrote, it is nonetheless "instructive to look at it like a game and to see where the points of likeness are."[63] One such instructive likeness was that "reasoning, like most games, is a social activity. That is, 'reasoning' is always an answer to the question, 'What are they doing.'"[64] As such, Rawls thought, it is carried out in accordance with certain generally accepted rules that create the possibility of moves and positions within the game. Rawls did not call these rules "constitutive," but his understanding of them in 1952 contained everything but the name. He listed several types of such rules, including those defining players and rules of etiquette, but he emphasized that the most important among them were principles, which "form the logical structure of reasoning … [by] govern[ing] what is to be accepted as a good reason, and rejected as a bad reason."[65]

Much like linguistic philosophers, Rawls began to view philosophical activity as an attempt to uncover the constitutive rules of ethical reasoning. This new view fit in well with his earlier positivist framework: Rawls could still view the philosopher as an analyst who tries to produce a theory, or an account of the constitutive principles that govern our reasoning. Thus, even if Rawls's explicit comparisons between ethics and a scientific inquiry disappeared and his new comparison was to logic, his adoption of the new beliefs did not necessitate changes in his earlier positivist framework.[66] Indeed, Rawls used the two together. Thus in his 1954 seminar on

---

[60] John Rawls, "Review," *The Philosophical Review* 60 (1951): 572-80; John Rawls, "On Values" (c. 1952), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7 Folder 9; John Rawls, "Ethics and Its Reasoning" (c. 1952), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 4; John Rawls, "Diseases of Ethical Reasoning" (c. 1952), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 14.

[61] Rawls, "On Explication"; John Rawls, "Oxford Notes: Spring 1953" (1953), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 10.

[62] Rawls, "Ethics and Its Reasoning"; Lecture 2, 1. All emphases in this dissertation are Rawls's.

[63] Rawls, "Ethics and Its Reasoning"; Lecture 2, 1.

[64] Rawls, "On Explication," 45.

[65] Rawls, "Ethics and Its Reasoning," Lecture 2, 2-5.

[66] Rawls, "Nature of Ethical Theory," 133, 7-8.

Christian ethics, he wrote that "to speak roughly the aim of the moral philosopher is to give a logical account of a good moral argument in much the same way that a logician attempts to give a logical account of a good deductive and a good inductive argument."[67] But in 1952, he stated that any particular ethical theory is a hypothesis about such criteria, and that a hypothesis of this kind is confirmed or refuted by reference to the actual ethical judgments.[68] Thus he continued to believe that the existence of ethical reasoning depended on one of such hypotheses being correct. Despite this change in overall outlook, then, Rawls still thought of ethics as an empirical inquiry, since it was "a question of empirical fact whether there is moral reasoning or not."[69] And that's what he set out to do: to determine "the sort of criteria which are used [in ethical reasoning] to distinguish good reasons from bad reasons."[70]

How to go about discovering these criteria was a different question. As Rawls acknowledged, unlike many games, ethical reasoning did not have explicit, written rules created by particular people at a particular time; in fact, the rules of reasoning were implicit and its rule-making body is everyone. As he wrote in "On Explication" [1953], "the striking thing about the constitution making body of reasoning is that its constitutional making body is <u>everybody</u>. It is part of <u>its</u> constitution that it has no official body."[71] Since there was no rule-making body in reasoning and no explicit rules, Rawls concluded that philosophers interested in discovering criteria for good reasons in ethics had to begin by analyzing the judgments of its informal constitution-making body: everybody. In that regard, he explicitly agreed with Toulmin's suggestion that we can discover criteria for good reasons by examining actual reasoning, or "various instances of the sort of reasoning in question and noticing how we actually distinguish between good and bad reasoning.'"[72]

Conceiving of ethical reasoning as a practice did not prevent Rawls from remaining a universalist. As we have seen in the previous chapter, he did not think that all instances of reasoning were useful for a philosopher tasked with giving an account of ethical reasoning.[73] Toulmin, he wrote, failed to specify just which instances of actual reasoning we should examine.[74] To correct this flaw in Toulmin's account, Rawls suggested that we restrict the range of relevant case studies. Although in his review Rawls did not explicitly limit relevant judgments to those that were considered and made only by reasonable persons, the reference to his "Decision Procedure for Ethics," where these notions are central, indicates precisely that.[75] Despite this restriction, Rawls remained as much a universalist as Toulmin: he thought that ethical judgments of all reasonable persons should be part of the subject matter of ethical theory. As he wrote in 1952, "the question whether there can <u>be</u> reasoning about a certain kind of question turns on whether or not that part of the constitutional body which sits in England agrees

[67] John Rawls, "Christian Ethics, Class at Cornell" (1954), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 5, "Some General Remarks on Christian Ethics: (Lectures 1 and 2)," 1.
[68] Rawls, "Ethics and Its Reasoning," Lecture 3, 1.
[69] Rawls, "On Explication," 77.
[70] Rawls, "Ethics and Its Reasoning," Lecture 1, 3.
[71] Rawls, "On Explication," 64.
[72] Rawls, "Review of Toulmin," 574.
[73] For description of the "reasonable person," see Rawls, "Remarks on Ethics," 5: 18-22. For considered judgments, see: Rawls, "Grounds of Ethical Knowledge," 47-60.
[74] Rawls, "Review of Toulmin," 574.
[75] Rawls, "Review of Toulmin," 574.

with that part of it which sits in India; or whether that part of it which sits in America agrees with that part of it which sits in Central Africa. Or if they do not now agree, can they upon mutual discussion and reflection come to an agreement on what the rules should be."[76] Thus, while Rawls disagreed with Toulmin about the kinds of ethical judgments that may serve as the subject matter of an ethical theory, he nonetheless agreed that an ethical theory is a universalist theory.

As Rawls's view of philosophy shifted to the study of the constitutive rules of reasoning, dilemmas associated with his earlier positivist framework fell to the side. Most notably, he was led to a new understanding of justification. Rawls's positivist analogy to science helped him only so far: it provided him with a model for objectivity in ethics but it also pushed him to posit "an objective moral factor" to the ethical situation itself in order to explain why all reasonable persons agreed in their judgments.[77] In 1950, Rawls escaped the push to moral realism by refusing to specify "how this common objective moral fact is to be interpreted."[78] By 1952 he no longer faced this problem because he accepted linguistic philosophy's belief that there is a point at which justification stops. As Rawls wrote in the review of *The Place of Reason in Ethics*, Toulmin was right to stress "the finite character of all reasoning, how in rational discussion it must be permissible to rest one's case at some point, how senseless it is to keep asking for a reason indefinitely."[79] This theme allowed Rawls to treat further demands to justify human agreement either as demands that do not arise or demands that cannot be meaningfully made. These arguments mark the emergence of a new notion of justification in Rawls's thought.

Rawls first developed the second argument. In his 1952 Princeton notes, which contain its earliest appearance, he introduced the notion of intuitive judgments, or judgments for which "no further reason can be given, or at least no one knows how to give one, and when no further reason seems necessary."[80] In 1954, calling intuitive judgments "inescapable," he related them to his previous notions of competent judges and considered judgments: intuitive judgments were such that "competent persons in their considered opinion find … [them] inescapable, and they can't imagine how an argument against them would go." [81] Rawls elaborated on this inability to raise sensible objections against intuitive judgments in his Cornell lectures. Inviting his students to imagine themselves sailing and spotting a floating life-boat without people in sight, he argued that we all would turn our boat in its direction. If a fellow sailor, unaware of the life-boat, asked us why we changed direction, we would give him a reason: namely, that we spotted a boat with no people in sight. His further question "why?" would, in this context, be difficult to grasp: "We can't clearly get straight what we would say if you still said: that's no reason for going off our course (in this case). We might think you were joking; but it wouldn't be funny."[82] Although this example contains no references to Toulmin, it is evident that Rawls's approach is drawn from the themes of linguistic philosophy: in his example, we provide a reason to a fellow sailor, but – much like in *The Place of Reason in Ethics* or in Wittgenstein's example of a schoolboy – the sailor raises an undefined question, "Why?," or "Yes, but why should we turn our boat?" Rawls

[76] Rawls, "On Explication," 64.
[77] Rawls, "Grounds of Ethical Knowledge," 47-8.
[78] Rawls, "Grounds of Ethical Knowledge," 278.
[79] Rawls, "Review of Toulmin," 574-5.
[80] Rawls, "Ethics and Its Reasoning," Lecture 3, 15.
[81] Rawls, "Christian Ethics," 15.
[82] Rawls, "Christian Ethics," 15.

emphasized that the sailor no longer provides reasons which we can consider: he does not, for instance, say that changing course in these circumstances is dangerous. As a result, we do not know what the fellow sailor finds objectionable in our decision. Thus, Rawls asked – at this stage of the argument almost rhetorically – "what can be proposed as an alternative statement [to our decision] and what would be the form of reason involved in this alternative statement?"[83] He did not answer this question. Instead, he concluded that the sailor's 'Why?' rested on doubts, but on doubts which did not stem from actual commitments. The underlying thought must have been that this abstract doubt does not stem any of our actual practices. For this reason, the fellow sailor could not propose an alternative course of action. Insofar as these doubts did not connect to the rest of our conceptual framework, they were not doubts which we could articulate and to which we could respond. Having reached this point, Rawls thought, reason giving can be allowed to stop.

In his "Oxford Notes," Rawls made a second, much broader, argument against the further request for reasons. It relied on another theme from Toulmin: the function of ethical reasoning. Toulmin claimed that the function of ethical reasoning was "to correlate our feelings and behavior in such a way as to make the fulfillment of everyone's aims and desires as far as possible compatible."[84] Rawls also started relying on the function of ethics: as he wrote in 1954 at Cornell, the function of arguments about justice in particular was to decide between competing claims.[85] Requests for further reasons arise, Rawls thought, only in special circumstances, namely, when people disagree. Thus, even if further reasons can be given, they need not be given as long as there is a general consensus on reasons already provided: "we only need to show that … [our account of the principles of justice is] such that a competent person is willing to admit that he stands on it without further reasons, whether or not further reasons can be given."[86] That this is so, Rawls wrote, is an important point about the concept of justification: "if there is general willingness to stand on [our account of principles] there is no (general) obligation to give any further reasons, for the obligation to give reasons only arises where there is not general agreement."[87] Insofar as the function of ethical reasoning was completed, there was nothing else for an ethical argument to do. Thus, justification, as Rawls now understood it, was a passing of the burden of proof: if an objection is made, there is an obligation to respond to it; once it is dealt with, the burden of proof is passed onto an imaginary objector, and one need not give further reasons.[88]

Summarizing this new conception of justification, Rawls now explicitly rejected his earlier views on the topic and dilemmas that accompanied them. If in graduate school he was repeatedly led to ask whether human agreement was justified because of an objective moral fact or whether the moral fact was made objective by the human agreement, in 1953 he rejected this formulation of the question. "This recognition and acceptance [by competent judges] isn't what makes the principles exist," he wrote; "for to talk this way is nonsensical."[89] In his Cornell

[83] Rawls, "Christian Ethics," 21.
[84] Toulmin, *Reason in Ethics*, 137.
[85] Rawls, "Christian Ethics," Some Questions, 4i.
[86] Rawls, "Oxford Notes," 13.
[87] Rawls, "Oxford Notes,"13.
[88] Rawls, "Christian Ethics," 20.
[89] Rawls, "Oxford Notes," "The Fundamental Problem of Ethical Theory: Are Principles Strong Enough?"

lectures a year later, Rawls made the point more forcefully, contrasting two different notions of justification: "We don't offer the general opinion of competent persons as a further reason for the judgment, or as justification of it, but we do offer it as justification for putting the burden upon him who would doubt."[90] From 1952 onwards, Rawls set aside justification as the giving of further reasons for human agreement, and the problems stemming from that framework ceased being problems for Rawls.

Philosophy and Politics: the Basic Structure, Justice as Fairness

Linguistic philosophy's notion of 'practice' prompted a second important development in Rawls's thought: it allowed him to draw limits of his inquiry to the major social institutions, or practices, of a society. While Rawls started drawing the boundaries of the concept of justice independently, his later steps were markedly influenced by Toulmin. Rawls turned to the problem of justice only in the early 1950s, after finishing his dissertation, which analyzed our judgments on the moral worth of character. In "Delimitation of the Problem of Justice," the latest in the series of writings not yet touched by the themes from linguistic philosophy, Rawls started his writings on the subject by drawing the boundaries of the concept of justice. "The problem," he wrote, "is to determine when acts cease to be morally indifferent, and when the question of their justness may be appropriately raised as an issue."[91] His idea was to restrict the application of the concept of justice to situations where important interests of at least two people conflicted substantially.[92] Although Rawls did mention the most important of such conflicts, namely, those involving "life, liberty and the pursuit of happiness," he did not restrict his analysis of the concept of justice to any particular conflict of interests.[93] By 1953-54, however, he had confined the subject of justice to the institutions and the constitutional structure of society.[94] This evolution took place in two steps, both inspired by the new conception of ethical reasoning as a human practice.

The first step was to restrict the application of "justice" to institutions or practices, which Rawls did in the 1952 lectures at Princeton. It is important to note – as Rawls did – that one had to provide a further argument for the move from understanding ethical reasoning as a practice to claiming that it was a second-order practice that regulates other practices.[95] One could as well, he thought, have analyzed the justice of individual acts.[96] Rawls's main argument for this step was influenced by Toulmin's account of justification, which distinguished between justifying individual acts and justifying practices or institutions. Toulmin thought that reason-giving in the former consisted essentially in appeals to principles; once the act was shown to be justifiable by

[90] John Rawls, "Justice as Fairness, Cornell Seminar 1953 Fall" (1953), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 11, 15.
[91] Rawls, "Delimitation of the Problem of Justice" (c. 1950), John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 9 Folder 13, 1.
[92] Rawls, "Delimitation of the Problem of Justice", 1-4.
[93] Rawls, "Delimitation of the Problem of Justice", 4.
[94] Rawls, "Justice as Fairness. Cornell Seminar," 3.
[95] Rawls, "Diseases of Ethical Reasoning," 34.
[96] Rawls, "Diseases of Ethical Reasoning," 38.

the rules of the institution, one could only criticize it by criticizing the institution itself.[97] In that regard, he treated justification of institutions as primary.

Toulmin's distinction struck Rawls as very useful, and, in fact, in 1955 he published "Two Concepts of Rules," suggesting a defense of utilitarianism based on this very distinction.[98] Meanwhile, in 1952 he used it to restrict the subject of justice. Actions, Rawls wrote, are justifiable by appeal to rules, most of which fall in the context of some institution.[99] To evaluate an action or a rule, we must, therefore, consider the relevant institution: "Thus we ask: do these rules, as directives to be followed, accomplish in the best possible manner, the purpose of the institution of which they are a part?"[100] Although Rawls did not explicitly conclude that, therefore, a philosopher in ethics should concentrate on the justice of institutions, given the reasoning so far this step seems natural to make. In his 1953-54 lectures at Cornell, Rawls experimented with a stronger argument: appealing to the use of the word 'just,' he claimed that we do not apply this label to particular actions but, rather, to the institutions of which they are part. For example, he wrote, breaches of legal rules have their own labels, such as "murder, theft, assault" and so on; but we do not say of these actions that they are unjust, restricting the use of this label to the corresponding institutions: "particular actions so covered are not said to be just or unjust; it is the institution itself which may be just or unjust."[101] Later in the same lectures, however, Rawls brought to light cases where particular actions are in fact called just or unjust; his eventual decision to restrict the subject of justice to practices must have therefore rested not on the argument that the label 'just' is not actually applied elsewhere, but – as he argued in *A Theory of Justice* – on the judgment that the justice of practices is a more important subject due to its long-term effects on people's prospects in life. In 1953-54 his reasoning may have been similar, but it was not stated in print.[102]

Rawls took the second step in the delineation of the concept of justice – selecting a range of practices to which the use of 'just' applied – already after his visit to Oxford. Doing so, Rawls developed the idea that justice is a form of fairness. The crux of the idea was that in the real world which we inhabit we cannot but play some games and partake in some practices. This is particularly true of practices defined by major social institutions, such as those regulating the constitutional structure of society. But if we can use the notion of fairness to discuss games, it is more appropriate to use the notion of justice to describe those particular games which we cannot but play. "Remember," Rawls told his students at Cornell in 1953-54, "it is one thing essential about the notion of a game that people have an <u>option</u> whether or not to play."[103] The notion of fairness, he continued, "has its home in games:" games are fair when the rules explaining how winning is possible are clear, when we win and lose by the rules.[104] "When we must play the game," however, a new notion – that of justice – enters.[105] Since, he argued, "we must play the

[97] Toulmin, *Reason in Ethics*, 146.
[98] John Rawls, "Two Concepts of Rules," *The Philosophical Review* 64 (1955): 3-32.
[99] Rawls, "Diseases of Ethical Reasoning," 6.
[100] Rawls, "Diseases of Ethical Reasoning," 6.
[101] Rawls, "Diseases of Ethical Reasoning," 27 (2).
[102] Rawls, *A Theory of Justice*, 10-11.
[103] Rawls, "Justice as Fairness. Cornell Seminar," "Reasoning and Games," 7.
[104] Rawls, "Justice as Fairness. Cornell Seminar," "Reasoning and Games," 8, 7.
[105] Rawls, "Justice as Fairness. Cornell Seminar," "Reasoning and Games," 8.

game which our social institutions impose upon us," "justice is here appropriate."[106] However, living in a society we cannot but participate in a wide range of practices; to narrow down this range Rawls introduced a further restriction on the concept of justice: that the stakes of these practices be high: "justice is what fairness becomes when the 'game' is compulsory and the stakes are high."[107] Thus, the name of Rawls's eventual conception of justice – "justice as fairness" – originated in the Wittgensteinian analogy of social practices to games, and was meant to restrict the analysis of justice roughly to those institutions in which we cannot but participate and the stakes of which are high.

Theory and Practice

If Rawls's transition to the new conception of justification and his delimitation of the subject of justice were relatively seamless, the themes from linguistic philosophy pushed him to much more problematic changes in his conception of ethical principles. Following positivists such as Ducasse, Rawls thought that it was possible to discover principles which underlie all of our judgments of justice. This view rested on two independent assumptions: first, that, if we thought about the issue carefully, we would *all* decide questions of justice for the same reasons, and second, that we would decide *all* questions of justice for the same reasons. I have been referring to the first assumption as that of universality, and I have named the second the assumption of the homogeneity of judgment. This second assumption rested on either of the two claims: that the relevant conceptual framework to decide questions of justice is the same in all situations in which such questions arise, or that the principles are reasons which are sufficient to determine our ethical judgments no matter what other relevant reasons are present. Resting on this homogeneity assumption, the principles were meant to serve as major premises in the deduction of particular judgments. As Rawls wrote in 1946, "we propose to construct a theory, to make deductions from it, and to test these deductions against the subject matter of ethical theory, namely, the actual moral judgments made by the class of people whose judgments constitute the reference of the theory."[108] Positivist political theory held the promise of guiding political practice by offering principles from which practical judgment could be deduced.

Linguistic philosophers such as Hampshire and Hart questioned the homogeneity assumption and, together with it, the ability to use ethical principles as premises for deduction. Rawls acknowledged the force of their objections already in 1951, and, in fact, used them against Toulmin in his review of *The Place of Reason in Ethics*, criticizing him for assuming that ethical judgment is homogenous. First, Rawls wrote expressly following Hart, ethical rules or principles are "defeasible": "certain standard exceptions are allowed for, and also openings are left for the entirely unexpected."[109] Second, he argued, the weight of reasons cannot be determined in advance, and so any one reason cannot be said to always prevail over all other relevant reasons: "[while] it is true that if there is a recognized rule then appeal to it always has *some* force … the

[106] Rawls, "Justice as Fairness. Cornell Seminar," "Reasoning and Games," 8.
[107] Rawls, "Justice as Fairness. Cornell Seminar," "Reasoning and Games," 1-3.
[108] Rawls, "Nature of Ethical Theory," 29.
[109] Rawls, "Review of Toulmin," 577-8.

force of the appeal varies from one kind of case to the next."[110] Summarizing his discontent with an account like Toulmin's, in his 1952 lectures Rawls explicitly sided with Hampshire, arguing that "a search for definitions or verbal equivalences is often done under the assumption that there must be some single sufficient reason from which one must always and necessarily base one's judgment; and further that this is a mistake."[111]

This clearly indicates a change in Rawls's view of how helpful political theory can be in political practice. The change is masked by the persistence of the term "theory," but the term now acquired a new meaning. Rawls now started understanding the principles as guides that highlight reasons relevant in political judgments and argued that no reason is always decisive. Therefore, he thought, principles should be understood as "logically loose" guides to judgment. As he put it on the margins of the 1952 lectures at Princeton, the principles act not as premises for deduction but as "bins, boxes of reasons."[112] "It is characteristic of moral arguments," he wrote explaining this new conception of ethical theory, "that the principles always constitute a form of good reasons; but the application of no single principle need be conclusive. There is no conclusion at all in the sense of there being a conclusion to deductive and inductive arguments."[113] Instead of looking at the principles as the premises for deduction, he wrote, we should view them as indicating "reasons supporting a certain course of action."[114] In the 1952 paper "On Values," Rawls incisively called the principles "rules of relevance," or "instructions as to what aspects of a situation are relevant."[115] The boxes will contain many reasons, but not all of them will be relevant in any one particular case. In addition, the weight of these reasons cannot, he thought, "be precisely determined" in practice.[116] So, instead of being a straightforward deduction from the principles given by an ethical theory, our decision "depends upon what other reasons there are and how the reasons taken as good support one another or fend one another off."[117] As Rawls wrote in a summary, "no valid account of ethical reasoning about justice can take the form of a cook-book code. It may be precise and clear, as I hope my account is; but it will remain logically loose: a way of patterning, arranging and testing for valid argument, but not a mechanical way of grinding them out."[118]

With this change in the conception of principles, the clear link between political theory and political action present during Rawls's graduate years was now giving way to a much more indeterminate connection. In 1950, he defended his principles as mediators in cases of disagreement: understanding them as premises for deduction, Rawls thought that principles were sufficient to determine the judgments of all reasonable men in all particular circumstances, and in the same direction. It seemed that the assumption of the homogeneity of our judgment was the key support for his belief in universality. Now, loosening the requirement of the homogeneity of

[110] Rawls, "Review of Toulmin," 578.
[111] Rawls, "Ethics and Its Reasoning," "Meaning of Good," 1. Rawls's criticism of Toulmin is not entirely fair – see Toulmin, *Reason in Ethics* 148, where he acknowledges that principles are not expected to be helpful in every situation that requires judgment.
[112] Rawls, "Diseases of Ethical Reasoning," "On the Application of Our Criterion in Concrete Cases," 5.
[113] Rawls, "On Explication," "The Fundamental Problem of Ethical Theory: Are Principles Strong Enough?", 2.
[114] Rawls, "On Explication," "The Fundamental Problem of Ethical Theory: Are Principles Strong Enough?" 2.
[115] Rawls, "On Values," 15.
[116] Rawls, "Diseases of Ethical Reasoning," 2.
[117] Rawls, "On Explication," "The Fundamental Problem of Ethical Theory: Are Principles Strong Enough?", 2.
[118] Rawls, "Diseases of Ethical Reasoning," 2.

our judgment, Rawls was in danger of harming both the usefulness of his theory to practice and his key belief in its universality. As the principles were "logically loose" guides to decision, it was no longer evident that the same ethical principles would determine the decisions of reasonable persons in the same direction. In fact, in the 1952 lectures he admitted for the first time that even people who reason appropriately may not agree on particular courses of action, as, given their varied backgrounds and experiences, they may assign different weights to the same reasons.[119]

To foreshadow questions which will arise because of this admission, it needs to be explained how potentially damaging this change in the conception of ethical principles is for Rawls's theory. Despite his modification of positivism, Rawls still thought that the objectivity of ethical reasoning rested on finding agreement in the judgments of reasonable men. As he treated this agreement as a stopping point in justification, it was a crucial assumption in his conception of philosophy. Now, however, Rawls acknowledged that reasonable persons' judgment may diverge and that agreement may therefore not obtain. Failing to find agreement in the judgments of reasonable men, Rawls would be forced into one of two alternatives: either to select some of the judgments of reasonable persons as better than others, or to provide a way of saying that all judgments, despite differences, are nonetheless in some essential respects the same. The first option would have required providing a standard for selecting between the judgments of reasonable men, a step that would have demanded a drastic change in Rawls's conception of philosophy. Perhaps for this reason, up until the first reviews of *A Theory of Justice*, this move never appeared to Rawls as a possible move. The second option, much more congenial to his approach, would, as we will see in the next chapter, be his choice in the late 1950s. Meanwhile, in 1953-54 he addressed the fact of disagreement explicitly, but he did not raise it as an actual threat to his understanding of philosophy or hesitate in his answer. He thought that the principles conceived as bins of reasons were strong enough to determine our judgments in the same direction. As he wrote in his Cornell lectures, "what makes [a principle a] principle is the way it is <u>used</u>: ie., to resolve difficulties and conflicts"; accordingly, when formulating principles "we would expect anyone to be able to assent to them, and then use these principles in such a way that they do straighten out many controversial particular matters."[120] He did not explain just how the principles would make our judgment converge in cases of disagreement, and this lack of concreteness will force him to come back to the question in later years.

Rawls at Oxford and Cornell: Towards the "Original Position"

It may appear ironic that Rawls had elaborated the core of his new conception of justification before going to Oxford, a hub of linguistic philosophy, and that, while at Oxford, he worked on tasks that required decision theory and formalism to which Oxford linguistic philosophers were very much opposed. In 1952, encouraged by a visiting Oxford philosopher J.O. Urmson, Rawls applied for, and won, a fellowship at Christ Church, Oxford. While at Oxford, he became personally acquainted with many of the aforementioned philosophers. At least once, when H.H. Price could not attend, he participated in a discussion with Ryle, Mabbott

[119] Rawls, "Diseases of Ethical Reasoning," 5.
[120] Rawls, "Justice as Fairness. Cornell Seminar," 22, 1.

and Kneale.[121] He conversed with Isaiah Berlin, frequently and at length, returning to his rooms in late-night hours.[122] These younger generation philosophers would become lifelong friends of Rawls. In the Spring of 1953, Rawls also attended Anscombe's lectures on Wittgenstein's *Philosophical Investigations*, which she translated and would publish later that year. These meetings did not have immediate effect: his two sets of notes written while at Oxford, "On Explication" [1953] and "Oxford Notes" [1953], as well as his 1953-54 lectures at Cornell, contain mainly continuations of themes developed during the two previous years. Rawls gathered these themes together into a consistent framework, and the rough shape of *A Theory of Justice* began to emerge, but no significant shift in Rawls's argument is noticeable in these years. Indeed, Rawls started developing models that departed from linguistic philosophy's characteristic contextualism.

But Rawls's turn to models of rational choice is intelligible against his earlier positivist framework. He understood linguistic philosophy in his own way, thinking that the use of thought experiments was entirely consistent with linguistic philosophy's emphasis on analyzing actual ethical judgments. His aim was still consistent with that of linguistic philosophy: to elaborate conceptual connections of 'justice' as the word is actually used – or, rather, as decisions about justice are actually made. Rawls already thought that Hampshire and Hart's understanding of ethical judgments could be reconciled with a modified account of ethical principles. He also thought that, in this modified form, principles of justice were strong enough to determine our judgment in the same direction. Thus, if all reasonable men decided in the same way, using models to analyze typical situations of justice must have looked to Rawls like only a more manageable way to carry out the analysis of justice. This analysis proceeded in two steps: in the first, the "pure case," Rawls made clear the general rules of ethical reasoning, while in the second, the "reasoning game," he considered what principles of justice these general rules would yield once applied to social institutions.

In his 1952 lectures at Princeton, Rawls took what turned out to be the first step in this direction: he developed the "pure case" experiment. The idea behind it was to approach the problem of justice with the "simplest sort of cases" that would nonetheless "throw light upon our problem."[123] Rawls did this in three ways. First, he described the thought experiment after the circumstances in which questions of justice arise. So, following his rough 1950 delineation of the concept of justice, he defined the pure case as one involving at least two people whose important and substantial interests conflicted.[124] Aspects not present in the situations of justice were excluded from the pure case. Second, he modeled the thought experiment after our judgments of justice. For example, he described the person in this pure case as rational in order to reflect our judgment that "persons are capable of deciding, and ought to have the <u>right</u> to decide what they want <…>; and [that] therefore there is never a question of forcing a good on a person, or forcing on him more than he puts in a claim for."[125] Third, Rawls excluded aspects of situations that complicate questions of justice. For example, he assumed that the goods to be distributed

[121] John Rawls, "Autobiographical Notes," [c. 1990] John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 42, Folder 12, 19-20.
[122] Rawls, "Autobiographical Notes," 19.
[123] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 1.
[124] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 2.
[125] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 2.

were given, and that their present distribution does not affect their future supply. Similarly, he took the number of persons and the time period as fixed.[126] Such simplifications, Rawls thought, would set aside important problems of justice, such as, in this case, justice between generations and, more generally, the impact of present redistribution on future growth. Once such complications were set aside, he thought, "we can then see what grounds there are left as relevant grounds."[127] He assumed, then, that the more complicated instances of justice introduce reasons peculiar to their issues, but that they all contain reasons of the simpler situations of justice.

Analysis of the described pure case consisted in making explicit our reasons for dividing fixed goods among people in this pure case of justice. Rawls thought that our judgments could be described by very general principles. For instance, he thought we would conclude that, barring relevant differences, every claim should be evaluated by the same principles, and no claim should be denied without a reason.[128] This principle of procedural equality – the term Rawls himself did not use at the time – was formal; however, when applied to the pure case it led to an important ethical conclusion: since there were no relevant differences between people in the pure case, the distribution of goods between them was to be equal. In this way, the pure case set up "our ideal (… [or] that state of affairs which we primarily want to bring about): a steady increase in freedom, in the development of human capacities believed to be goods, in the standard of living, all of which evenly passed around to all members of society."[129]

This standard of baseline equality served as an ideal by which to judge more complicated instances of justice in our actual societies. In these latter, Rawls thought, the pure case of justice never obtains and departures from the standard are expected: "I accept inequalities in any social context as inevitable, as necessary."[130] In fact, he thought, inequalities are desirable, as we buy progress at their expense.[131] The proper question about inequalities was not, then, whether they were justified, but which ones were justified.[132] Rawls's criterion for such cases was that of reasonable preference by all: institutions were justified as long as the inequalities they engendered were "functional, or effective, in increasing the amount produced at such a rate that it is reasonable for each man to prefer the benefits of the expected increase rather than to take the benefits of equal distribution now."[133] Just what sort of inequalities these would be Rawls did not specify; "this we cannot say," he wrote, "until we look at the facts," including the principles of economics.[134]

In the later years, Rawls developed a new thought experiment to reflect the conclusions reached in the 1952 lectures. Between 1952 and 1954 Rawls's description of this experiment was sparse but its aim apparent: to describe the imaginary situation of choice in such a way that it reflects the circumstances of justice and the conclusions reached in the "pure case" of justice. Although the idea was not well developed, this new choice situation was the beginning of the

[126] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 2-3.
[127] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 2.
[128] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 4.
[129] Rawls, "Diseases of Ethical Reasoning," Lecture 8, 2.
[130] Rawls, "Diseases of Ethical Reasoning," "On the Application of Our Criterion to Concrete Cases," 2.
[131] Rawls, "Diseases of Ethical Reasoning," "On the Application of Our Criterion to Concrete Cases," 1.
[132] Rawls, "Diseases of Ethical Reasoning," "On the Application of Our Criterion to Concrete Cases," 2.
[133] Rawls, "Diseases of Ethical Reasoning," "On the Application of Our Criterion to Concrete Cases," 1.
[134] Rawls, "Diseases of Ethical Reasoning," "On the Application of Our Criterion to Concrete Cases," 3.

hypothetical experiment that, in *A Theory of Justice*, would become known as the "original position." Rawls's description of this new choice situation was heavily influenced by the developing decision theory and its foundational works, such as John von Neumann and Oskar Morgenstern's *The Theory of Games and Economic Behavior*, which Rawls had studied while at Princeton.[135]

As it is clear from Rawls's later remarks, he was impressed by decision theory's ability to deduce conclusions from its assumptions. Yet, he started incorporating it in his work only in 1953, when he saw its fitting connection to his new analogy of reasoning to games. Games, Rawls wrote, are usually "decidable:" in typical circumstances, given the rules of the game, a winner can be determined.[136] Decision theory was capable of doing precisely that: it gave the problem a definite solution. In 1960, describing an ideal to which his own derivation of the principles of justice aspired, Rawls brought forth an example of a cake and two rational egoists. Given the task of cutting the cake with the aim of getting as much of it as possible, rational egoist A, knowing that a fellow rational egoist B has the same desire and the right to choose first, would always cut the cake into equal parts. That the cake would be so divided is not, Rawls emphasized, a psychological conjecture, or a hypothesis, but a conclusion that logically follows from the premises.[137] In reverse, if we, composers of the game, want the cake to be divided equally, we will describe the choice situation just as above. Similarly, if we want to see the implication of our standards, we will describe the choice situation according to these standards and see how the two egoists cut up the cake.

By 1953, Rawls had come up with just such an idea: to create a situation of choice, define it in a way consistent with our judgments about what is just and unjust, and see what definite results it would yield. In his Oxford notes, Rawls gave the first rough formulation of the new thought experiment: "the strategy is to design the game in such a way that these agreed criteria get forced on players; and this is done by making it in players' interest to opt for what they think just."[138] The plan was to have rational egoists choose the principles that will regulate the social institutions, but compel them to be reasonable in their choice.[139] Description of this situation of choice changed: in 1953, Rawls wanted to make the rational egoists propose principles of justice independently of each other, and to have these proposals be moderated by an official body.[140] The details need not concern us here, but the idea was that, not knowing which principles the official body will select, the parties will propose principles advantageous to themselves and fair to others.[141] In the 1953-54 lectures at Cornell, when the position of the lowest representative position in society was introduced, the choice situation started resembling that of cutting the cake: "A picture of how to make a rational egoist design a just society: let him design it and give

---

[135] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press, 1944).
[136] Rawls, "Justice as Fairness. Cornell Seminar," 7.
[137] John Rawls, "Political Philosophy 171" (1960), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35 Folder 12, Lecture IX, 5ii.
[138] Rawls, "Oxford Notes, 1953," "The Fundamental Problem of Ethical Theory," 12
[139] Rawls, "Justice as Fairness. Cornell Seminar," "The Reasoning Game," 4.
[140] Rawls, "Oxford Notes, 1953," 49. See Rawls's description of this early version of the "original position" in his interview with *The Harvard Review of Philosophy*: Aybar S., Harlan J. and Lee W., "John Rawls: For the Record" *The Harvard Review of Philosophy* (Spring 1991): 39-40.
[141] Rawls, "Oxford Notes, 1953," 50.

his worst enemy the option of assigning him his place in it."[142] Rawls was not explicit about the reasons behind this condition, but they are clear from the way he described the chosen principles of justice in the same lectures: "we might summarize [the principles], and so the notion of justice, by saying that they requir[e] of the various institutions of society that they start from the position of assuring equal and maximum freedom to every one and depart therefrom only in such a way as to make every man better off in the long run."[143] In short, the condition of the lowest position in society is required to make sure that inequalities are justified even to those who are most disadvantaged by them.

By 1958, when he would publish "Justice as Fairness," Rawls would have already arrived at the two principles of justice. They would reflect conclusions he reached in 1952 and 1953-54: the first principle would grant equal liberty to all those participating in a practice or affected by it, while the second would declare all inequalities arbitrary unless "it is reasonable to expect that they will work out for everyone's advantage" and provided that offices are open to all.[144] Throughout the years, Rawls would add to the description of the choice situation to ensure that the rational deliberators do choose the aforementioned principles. He would eventually give it a name by which we all know it – the "original position." Yet the main idea behind the thought experiment – to draw out the implications of our judgments of justice – would remain the same as presented in the 1953-54 lectures at Cornell.

Conclusion

From the late 1930s onward, the emerging linguistic philosophy was transforming the landscape of Anglophone analytic philosophy. This influence was particularly strong in the late 1940s and the early 1950s, when a new generation of linguistic philosophers started publishing their works. John Rawls was among the thinkers whose approach to philosophy underwent significant changes after encountering these works. His positivist conception of justification, reluctantly reliant on "objective factors" in the ethical situation to explain the convergence of reasonable persons' judgments, now gave way to a vision in which reason-giving stopped after all the interlocutors were satisfied with the conclusions and which rejected requests to give further foundations for this agreement. Reliance on human agreement also set Rawls's main goal: to show that it actually obtains. Thus in the early 1950s, he started constructing thought experiments meant to analyze the logic of ethical reasoning. His overall understanding of ethical reasoning as a human practice led to the restriction of the subject matter of justice to the main social institutions, or what will later become known as the basic structure of society. Finally, in response to arguments of Hampshire and Hart, Rawls modified his conception of ethical principles, without, however, abandoning his aim to build a theory organized around principles.

These new themes in Rawls's philosophy are responsible for many criticisms he would receive after the publication of *A Theory of Justice*; his allegedly quietist reliance on human agreement is the main among them. The linguistic themes would also shape his questions

[142] Rawls, "Justice as Fairness. Cornell Seminar," 9.
[143] Rawls, "Justice as Fairness. Cornell Seminar," 14.
[144] John Rawls, "Justice as Fairness", in Rawls, *Collected Papers*, 48.

between 1956 and 1964, when he would get further acquainted with the linguistic tradition and its concerns about the nature of necessity. Influenced by these encounters, in 1958 Rawls would offer a groundbreaking seminar on moral feelings at Cornell, in which, like the linguistic philosophers, he would attempt to draw necessary connections between moral emotions and moral principles. This period would also mark Rawls's solution to the problem of disagreement in the judgments of reasonable men: introducing a broad notion of 'sameness despite difference,' he would provide a way in which the same principles can explain the divergent judgments of reasonable persons.

5

Theory as a Guiding Vision

The eight years between 1954 and 1962 mark Rawls's most creative period and also one the results of which are least visible in *A Theory of Justice*. These years are Rawls's most Wittgensteinian years and ones in which his work is most clearly targeted against the dominant logical positivist positions.[1] The folk narrative which sees Rawls presenting philosophy with a post-positivist vision has most purchase in these years, as he offers his own alternatives to the positivist conceptions of necessity and moral emotions. It is also during these years that the main distinctions in Rawls's political philosophy form and the contours of his philosophical naturalism – his attempt to connect moral judgments with other human practices and capacities – are laid out.

These developments in Rawls's thought are related to the broader post-analytic turn in Anglophone philosophy and the main dilemmas of Wittgensteinian thinkers. In ethics, the post-analytic turn centered on discussions about emotivism. Understood as an extension of the analytic-synthetic distinction, one which denied any necessary connection between reasons and the so-called moral attitudes, emotivism was seen as an obstacle that had to be displaced before new conceptions of philosophy gained credence. Like all traditions of thought, emotivism was constantly evolving, and its new expressions, such as Richard M. Hare's *The Language of Morals*, were now more defensible.[2] This struggle between emotivism and the Wittgensteinian strands was condensed to the question about the relation between a word and the criteria for its application. Relying on their analytic-synthetic distinction, emotivists claimed that any attempt to tie the meaning of a word to its criteria would have to be a stipulative definition. Further, they argued, any such attempt would commit one to the absurd claim that 'good' had as many meanings as its applications. Thus Wittgensteinians faced the dilemma of specifying the nature of necessity that avoided the charge of arbitrary stipulation and the absurd implication that no two applications of the word can be the same.

By the mid-1950s, Rawls began to see this concern as his own. As is evidenced by his teaching notes on moral feelings and political philosophy, he sought to draw necessary connections between justice and concepts such as liberty, equality, the common good, as well as shame, sympathy, and pride. These new concerns are responsible for Rawls's changed way of understanding agreement among reasonable persons: appealing to the notion of the family

[1] Rawls studied Wittgenstein's work with intensity, going as far as to create a lexicon of Wittgenstein's terms. See his notes in John Rawls, "Wittgenstein investigation, lexicon" [1953]. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 60; John Rawls, "Wittgenstein Criteria" [1953]. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 8, and John Rawls, "Wittgenstein Investigations" [1953]. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 2.

[2] Richard Hare, *Language of Morals* (Oxford: Clarendon Press, 1952). For other examples known to Rawls see Axel Hagerstrom, *Inquiries into the Nature of Law and Morals*, ed. Karl Olivecrona (Stockholm: Almqvist & Wiksell, 1953); Jonathan Harrison "When is a Principle a Moral Principle?" (Symposium with Philippa Foot). *Proceedings of the Aristotelian Society*, Supp. Vol. 28 (1954): 111-34.

likeness, he now definitely gave up on the idea that reasonable persons would always agree on particular political judgments. This step led him to believe that philosophy is not always helpful in solving practical problems and that the most it can often do is guide our judgments in a "general direction."[3] After the publication of *A Theory of Justice*, this conclusion would elicit criticisms from various camps, including interpretivists, democratic theorists and in particular Marxists.[4] Wittgensteinian concerns are also responsible for the outlines of Rawls's naturalism. Wittgenstein's writings shaped Rawls's naturalism in two ways: he started viewing morality and justice as an outgrowth of natural feelings and began explaining the possibility of agreement among reasonable persons as a result of a shared background of natural feelings. All moral views consistent with the naturalist premises, he thought, will share various – and significant – family likenesses.

In this chapter, I want to show Rawls's participation in the debates of the Wittgensteinian tradition and his attempt to use Wittgenstein's techniques in his own inquiries. I first describe debates about the nature of necessity and show how, by around 1956, the two concerns of Wittgensteinians became Rawls's concerns. I focus on his conception of necessity as a conceptual connection which cannot be abrogated without a drastic change in the rest of our beliefs. In the second and third parts, I show how these dilemmas determine the character of Rawls's inquiries in his courses on moral feelings and political philosophy and how they lead to new understandings of theory and the place of moral judgments in human life.

Conceptual Necessity

Wittgenstein and philosophers influenced by him offered a new notion of necessity, one that emphasized conceptual connections the breaking of which made one's position unintelligible. This notion of unintelligibility made certain connections necessary and others impossible; concepts could not be uprooted from some of their contexts and could not be placed in some others. In ethics, Hampshire and Hart argued that one could not arrive at one definition of 'good,' as the meaning of 'good' varied from context to context. They both therefore thought that the meaning of 'good' was necessarily tied to its criteria. Wittgensteinians argued that, despite the differences in criteria, one could nonetheless say that some of the uses of 'good' were the same. They did so by appealing to the notion of family resemblance. These two concerns within the Wittgensteinian family – the nature of necessity and sameness of meaning – would become objects of contention and therefore their main dilemmas. Objections to their conception of necessity were most notably raised by two Oxford philosophers – Richard Mervyn Hare [1919-2002] and J.O. Urmson. As Hare was a critic of the Wittgensteinian approach, whereas Urmson in many ways favored it, these objections show that the two dilemmas were perceived as such both by critics and insiders.

---

[3] John Rawls, "Political Philosophy 171, 1962" John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folders 8-13, "Some notes on the Use of Political Philosophy," 1i.

[4] Michael Walzer, *Spheres of Justice: A Defense of Pluralism and Equality* (New York: Basic Books, 1983), 79, 82; Amy Gutmann and Dennis Thompson. *Democracy and Disagreement* (Cambridge, MA: Harvard University Press, 1996), 35-40. For Marxist criticisms, see the special issue on the reception of Rawls in Europe: Cécile Laborde, ed. "Rawls in Europe," *European Journal of Political Theory* 1, Special Issue (2002).

Hare's argument against contemporary accounts of necessity would pose a dilemma to thinkers like Rawls who attempted to draw more-than-contingent connections between 'justice' and its criteria, or principles. The key problem with contemporary accounts of necessity, Hare wrote in *The Language of Morals*, was that the allegedly necessary connections could always be reasonably questioned. Retrieving G.E. Moore's 'open question' argument and bringing back the label of "naturalism," Hare claimed that the meaning of 'good' could not be tied to its criteria either in general ('good' means 'a, b, c') or in particular instances of its use (*this* 'good' means 'a, b, c'). He argued against the general identification of 'good' with its criteria by showing that one could always ask whether the criteria given as part of the definition are themselves good, and against the particular identifications by questioning, for instance, whether an auger that 'bores holes well' is actually a good auger.[5] While virtually all Wittgensteinians accepted that 'good' did not have a criterion common to all contexts of its use, Hare's argument against particular identifications of 'good' with its criteria threatened the core of their position: if Hare was right that 'good' is independent of its criteria, then any connection drawn between the two was contingent.

Equally troubling was Hare's implication that the only way to draw a necessary connection between 'good' and its criteria was to do so by stipulative definition. Hare thought that we could, if we wanted, define a 'good painting' as 'a painting that the members of the Royal Academy admire,' thereby making the connection necessary in virtue of the stipulated meaning equivalence alone. Yet, this move would be uninteresting and misleading, as it overlooks the main function of 'good' – to guide choices. Hare used an argument by substitution to demonstrate this crucial mistake: defining 'a good painting' as 'a painting that the members of the Royal Academy admire', he rewrote a commendatory statement 'the members of the Royal Academy admire good pictures' into the trivial 'the members of the Royal Academy admire pictures that they admire.'[6] As the choice of the substitution argument shows, Hare took the naturalist argument to be one of equivalence, made by definition. Not only did Hare not mention any other plausible account of necessity, but he also claimed that the choice of moral principles is a decision that is never determined by our current beliefs.[7] Although he claimed that such a decision was not arbitrary, he did not think that it was necessary either. All these arguments suggested that Hare saw analytic necessity, or necessity by definition, as the only plausible account of necessity. Thus, he wanted to impose a two-edged dilemma on those pursuing the Wittgensteinian – or any other – approach to necessity: either the connection between 'good' and its criteria is contingent, or it is necessary but stipulative. Either way, the proposed allegedly necessary conclusions would be philosophically uninteresting.

J.O. Urmson's argument in "On Grading" was more exploratory than straightforwardly critical, but the problems he raised about the nature of necessity were those identified by Hare. Like Hare, Urmson presented a limited array of the types of connections: those that were analytically necessary, synthetically necessary and contingent.[8] He also claimed that naturalism took 'good' to be a mere shorthand for its various criteria, an instance of analytic necessity, or necessity by definition. Like Hare, Urmson thought that naturalism overlooked the fact that

[5] Hare, *Language of Morals*, 84-91, 103-108.
[6] Hare, *Language of Morals*, 84-85.
[7] Hare, *Language of Morals*, 65-69.
[8] Urmson, "On Grading," 154-56.

'good' was a grading label, or a label used to sort objects or activities into categories.[9] Unlike Hare, however, Urmson acknowledged that naturalism did contain some insights, in particular in stressing what he called the "close connection" between good and its criteria.[10] In consequence of this connection, Urmson thought, one could not always ask whether something that met the criteria for 'good' was actually good: in some circumstances, this question itself would be odd. Urmson also pointed out the second difficulty with accounts of necessity like those of the Wittgensteinian tradition: given that the criteria of 'good' vary from context to context and that no common criterion to all uses of good can be found, any attempt to define 'good' in terms of criteria would entail an absurd conclusion that 'good' had as many meanings as criteria.[11] Urmson thought, then, that Wittgenstein's argument of family resemblance and his motto that meaning is use threatened the possibility of any account of necessity. For that reason, he did not go further than to note the "close connection" between 'good' and its criteria. These two problems – specifying the nature of necessity that is not an arbitrary definition and elaborating a notion of 'sameness' which would allow for difference in criteria but be able to exclude contingent connections – would become important dilemmas for the Wittgensteinian tradition.

These dilemmas, posed well by Hare and Urmson but equally evident to Wittgensteinians who were not familiar with these writings, soon prompted response. In the general field of philosophy writers like Norman Malcolm and Rogers Albritton [1923-2002], both Rawls's colleagues at Cornell, attempted to explain Wittgenstein's distinction between criteria and symptoms, which stood, respectively, for necessary and contingent relations. In ethics, Peter Geach [b.1916] defended the identification of the meaning of 'good' with its criteria by rejecting Hare's false dilemma of necessary-but-stipulative or contingent and elaborating his own notion of necessity. He brought forth a new conception of necessity which connected definitions of words to human activities in which they played a part. A definition implied a certain human behavior: "It belongs to the ratio of 'want,' 'choose,' 'good,' and 'bad,' that, normally, and other things being equal, a man who wants an A will choose an A that he thinks good and will not choose an A that he thinks bad."[12]

Attempting to provide an analysis of the concept of justice, Rawls was interested in the questions raised by what he called the "Urmson-Hare thesis."[13] Until about 1956, Rawls's notes do not reveal direct discussions about the nature of necessity, or descriptions of his own project as one that draws some kind of necessary connections between 'justice' and its principles, or criteria. Yet, this is certainly what he did, claiming that his analysis of justice would explain actual judgments of reasonable persons and thereby draw a necessary link between justice and reasons given when judging something just. He therefore needed to clarify the nature of necessity on which his arguments relied. Similarly, he acknowledged the force of the family resemblance argument and yet wanted to say that justice is connected to roughly the same concepts in all contexts of its use, once this use is restricted to the basic institutions of a society.

[9] Urmson, "On Grading," 161.
[10] Urmson, "On Grading" 155, 156, 157.
[11] Urmson, "On Grading" 159, 162.
[12] Peter Geach, "Good and Evil," *Analysis* 17 (1956): 38.
[13] John Rawls, "Rational Choice and the Concept of Goodness" (1956), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 3, 2: 1i.

He therefore needed a way to say that reasons connected with the concept of justice are similar despite their differences.

These questions became Rawls's questions by 1956. To formulate them, Rawls contrasted the Urmson-Hare thesis with Geach's view. According to him, the former argued that "while good has one meaning, the criteria for something's being correctly said to be good vary from type of object to type of object," while the latter denied that there is such a thing as "being just good or bad," instead there is only "a good or bad so-and-so."[14] With these summaries of arguments in front of him, Rawls asked: "What criteria have we for deciding how many senses a word has? How do we want to distinguish between senses?"[15] To decide, Rawls made explicit the implications of insisting on either view. "If we say with Urmson that 'good' has one sense," he wrote, "the point in saying this would be to emphasize the use of 'good' in grading…. 'Good' has a central place in the language game which constitutes a part of deliberation and choice."[16] Geach's view, on the other hand, pointed to the necessary connection between 'good' and its criteria: "if we say that the sense of good varies with the criteria we are pointing out that criteria are not simply symptoms of some … other thing which is good directly, or goodness itself, or itself a criterion of good (that is, the connection is not simply contingent)…."[17] While Rawls agreed with Geach's emphasis on necessity, he also thought that we would not want to say that 'good' has as many meanings as objects or activities to which it applies: "it would conceal what is common in these uses, and make the use of the same word a queer linguistic fact."[18] Thus, two concerns emerge from Rawls's notes on the disagreement between Hare, Urmson and Geach: first, the need to establish a necessary connection between concepts that is not analytic, and second, the need to elaborate a criterion for sameness despite difference. These concerns would guide Rawls's writing between 1956 and 1958.

In the notes of the period, we see Rawls defining necessary connections as those that we could not abrogate without significant changes in other concepts that describe our essential activities. He thought it helpful to contrast this account of necessity with analytic necessity as it was expressed in Charles Arthur Campbell's "Moral and nonmoral values."[19] Summarizing Campbell's article, Rawls noted that it was correct to point out that "the principles of rational choice are necessarily connected with 'is good'; necessarily they are its criteria, or a part thereof."[20] Consequently, he concluded that Hare's open question argument and its implication of contingent relations was flawed. Yet, Rawls continued, Campbell was mistaken to claim that the necessary connection was one by definition: "[rejection of the "open question" argument] does not commit us to a definition in any normal sense."[21] Instead, he wrote, necessary connections are those that cannot be severed without changing other important parts of our conceptual framework. Thus, discussing the principles of rational choice (which, for Rawls, was the equivalent of 'good'), he argued that these principles rest "on the totality of concepts with

[14] Rawls, "Rational Choice and the Concept of Goodness," 2: 1i ; Rawls, "Rational Choice and the Concept of Goodness," 6: 1i-1ii.
[15] Rawls, "Rational Choice and the Concept of Goodness," 9: 6ii.
[16] Rawls, "Rational Choice and the Concept of Goodness," 9: 9ii.
[17] Rawls, "Rational Choice and the Concept of Goodness," 9: 9ii.
[18] Rawls, "Rational Choice and the Concept of Goodness," 9: 6ii.
[19] Charles A. Campbell, "Moral and Nonmoral Values," *Mind* 44 (1935), 273-299.
[20] Rawls, "Rational Choice and the Concept of Goodness," 10: 3i-3ii.
[21] Rawls, "Essay V," 4: 7i.

their logical connections which we use to express ourselves when we make decisions, give reasons, state our wants and desires, etc."[22] Here he appealed to the notion of intelligibility described in Chapter Four, and argued that we would not be able to forgo the connections with crucial concepts like "wants" and "desires" without becoming unintelligible: "anyone who knowingly failed to act in accordance with these principles would be said by us to be acting irrationally; and we could not understand (could not make sense of) his conduct at all unless we saw that he was swayed by momentary impulse or sudden passion."[23] Because of this connection to crucial concepts of choice and desire, and because the lack of such connections would make us unintelligible, Rawls thought it wrong to say that 'good' was defined by an act of a theorist: "to say this suggests that we have simply stipulated something about one word, or one concept, whereas the connections reach out to the whole group of concepts related to choosing."[24]

This understanding of necessity would influence the direction of Rawls's lectures on moral feelings and political philosophy. He would try to draw just such necessary connections between moral and natural emotions as well as between justice and the related concepts of liberty, equality, and the common good. Doing so would help him see morality as a natural phenomenon and lead him to reformulate his understanding of how helpful political theory may be in guiding political practice.

Moral Feelings

In 1958, Rawls offered a seminar on moral feelings. As he was well aware, his treatment of the topic was novel. Emergence of this new approach was part of the larger post-analytic turn, which, as in the broader debates about the nature of necessity, was driven by the Wittgensteinian tradition. Bringing the fight against emotivists to the field of moral feelings was appropriate given the core emotivist claim that the connection between moral reasons and moral emotions was contingent. Many, including Rawls, understood the emotivists to argue that no moral feeling was necessarily tied to any particular moral reason, and that moral feelings could be developed toward any moral reason.[25] Thus thinkers belonging to the broad Wittgensteinian tradition and writing on moral feelings, notably John Niemeyer Findlay [1903-1987] and Philippa Foot [1920-2010], argued against emotivism by drawing necessary connections between moral feelings and moral reasons, and by criticizing the emotivist conception of moral feelings – what Findlay called the "inner-quality view" – more directly.[26] Rawls joined in these criticisms, acknowledging debts to the mentioned thinkers, and borrowing the broad approach to the topic from Wittgenstein.

Themes from Wittgenstein's philosophy and the writings of Wittgensteinians exercised a shaping influence on Rawls's naturalism in two ways. Most broadly, Wittgenstein's approach to

[22] Rawls, "Rational Choice and the Concept of Goodness," 16: 5i-6i.

[23] Rawls, "Rational Choice and the Concept of Goodness," 16: 5i.
[24] Rawls, "Rational Choice and the Concept of Goodness," 16: 6i.
[25] John Rawls, "Moral Feelings, 1960," John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 1, "Seminar I," 2ii; John Rawls, "Moral Feeling I" (1958), John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 34, Folder 19, "Summary," 1i.
[26] John N. Findlay, "The Justification of Attitudes," *Mind* 63 (1954): 149.

philosophical questions led Rawls to analyze the concept of morality – or the fact of having moral views – by placing it in the wider background of characteristically human activities. In particular, Wittgenstein's discussion of pain suggested to Rawls that moral views can be analyzed by looking at the moral emotions with which they are connected. This argument carried the conclusion that a person's moral views and sense of justice are intelligible as outgrowths of natural human emotions. Second, Wittgenstein's concepts of the "form of life" and "family likeness" provided Rawls with reasons to expect agreement in the considered judgments of reasonable persons. Interpreting the "form of life" in a restrictive way, Rawls thought that all feasible moral views – moral views that can be connected with natural attitudes – will share a family likeness. So, unlike Wittgenstein himself who used "family likeness" to argue against the essentialism of analytic philosophy, Rawls placed the emphasis not on the difference but on the likeness of moral views. This helped him to both acknowledge the pluralism of moral views and maintain his hope to discover commonalities shared by them all.

Wittgenstein's influence was most visible in Rawls's attempt to place morality in the background of feelings which made it intelligible. "With Wittgenstein I shall assume," Rawls told his students in the "Moral Feelings" seminar in 1958, "that <u>having a concept</u> is essentially mastering the use of a word in its proper background of thought and feeling."[27] In order to explicate a concept, Rawls continued, one examines "characteristic uses of the word … and the associated family of expressions," how someone is taught the word, the types of feelings it calls forth, the way such feelings are resolved, as well as the behavioral manifestations accompanying the feelings.[28] In short, Rawls's guiding principle was to look at "having morality as a form of life, or as an <u>aspect</u> of a form of life" and to treat morality "as a whole, as a natural phenomenon, as a complex of thought, feeling and action continuous with other aspects of human life."[29] His goal was not so much to argue against emotivism – despite long discussions of the topic in the seminar – but, more broadly, to outline the connections between a moral view and the complex of feelings that makes this view intelligible. These connections would allow him to argue that claims about natural human capacities have important – if nonetheless weak – implications for ethics.

At the center of his naturalism was the claim that human beings have a psychological tendency to recognize another human being as a person and not as an object, that is, a being "who has wants and interests, who experiences emotions like fear, grief; etc;… [who] is rational, is able to deliberate and decide; is able to state intentions and has memories."[30] His goal in this seminar was to explore how this recognition of others as persons expresses itself in the form of natural and moral feelings (and consequently moral reasons). It is noteworthy that Rawls attributed the origin of his emphasis on the recognition of another as a person to Wittgenstein. Summarizing this argument in "Justice as Fairness," Rawls wrote that his idea that "the response of compassion, under appropriate circumstances, is part of the criterion for whether or not a person understands what 'pain' means is, I think, in the *Philosophical Investigations*." His own argument, Rawls thought, was "simply an extension of this idea."[31] Although this attribution is

[27] Rawls, "Moral Feeling 1958," "Procedure," 1i.
[28] Rawls, "Moral Feeling 1958," "Procedure," 1i-2i.
[29] Rawls, "Moral Feeling 1958," Topic VII, 1i.
[30] Rawls, "Moral Feeling 1958," Topic IX: Compassion, 5i.
[31] John Rawls, "Justice as Fairness," in Rawls, *Collected Papers*, 62f.

overly generous – recognizing another as a person was a key feature of Rawls's undergraduate thesis, *Meaning of Sin and Faith* – it nonetheless shows that Rawls saw Wittgenstein's ideas as supportive of his own work and inspiring it in new directions.[32] Thus Wittgenstein's remark, "Pity, one may say, is a form of conviction that someone else is in pain," and his argument that the word "pain" is applied to human beings but not to dolls or stones led Rawls to develop his own version of naturalism.[33]

Explicitly using Wittgensteinian terms, Rawls argued that the recognition of another as a person logically entails acting in certain ways: "it is part of the criterion for one person's recognizing another as a person that he act in a certain way towards him; just as it is part of the criterion for a person's recognizing the difference between colors that he respond to them differently and show in his acts an awareness of their different relations."[34] In particular, he thought, anyone recognizing another as a person had to display natural and moral feelings: "My hope is that having a morality implies (in some way) having the natural feelings, and that having the natural feelings implies the recognition of persons as persons."[35] Again, Rawls understood natural and moral feelings very broadly. Unlike natural feelings, such as pride, joy, grief, anger and love, he explained, moral feelings like shame, guilt and remorse "presuppose moral standards of some kind."[36] Moral feelings, he clarified, are "explained and accounted for by reference to moral concepts," such as the principles of justice, whereas moral concepts may occur in the explanation of natural feelings, but need not do so.[37] Natural feelings were 'natural' in another sense as well: given normal, or typical, conditions of human life in which affection was present, it was expected that natural feelings like affection, love, and joy would develop. Thus, Rawls wrote, "this development [of natural emotions] is perfectly natural: that is, their affection and liking for companionship will occur if it is given only that minimum of environmental invitation found in the simplest conditions of group life, and so under the natural and normal conditions under which men have lived. The liking and capacity for friendship and affection is natural in that it develops under normal conditions."[38] Consistent with the larger purpose of showing that anyone – children and rare medical cases excluded – who failed to exhibit moral and natural emotions is an unintelligible being, Rawls understood both types of feelings very broadly and did not tie them to any specific conception of morality.

While Rawls spent most of the seminars drawing connections between particular moral feelings and the attendant concepts, he insisted on reminding the students that he did so with a larger purpose in mind: showing that there is a necessary connection between moral and natural feelings. Rawls insisted that the connections he drew were necessary, or "logical."[39] The intended contrast was with contingent connections; thus, he insisted that a person who has

[32] Rawls, *Meaning of Sin and Faith*, 110-121.
[33] Wittgenstein, *Philosophical Investigations*, §287 and §§281-287.
[34] Rawls, "Essay V," 8: 2i.
[35] Rawls, "Moral Feeling 1958," Topic III, 1i-1ii.
[36] Rawls, "Moral Feeling 1958," Topic IX, 4i, 5i; Rawls, "Moral Feelings 1960," Seminar III, 4ii; Rawls, "Moral Feeling 1958," "Procedure," 3i.
[37] Rawls, "Moral Feeling 1958," Topic II, 1ii-2ii.
[38] Rawls, "Essay V," 6: 1i.
[39] Rawls, "Moral Feeling 1958," Topic I, 4i.

natural attitudes would "necessarily exhibit … certain forms of moral behavior."[40] Recognition of another as a person, he argued, entailed both moral and natural feelings.

He thought it useful to distinguish two directions of connection between moral and natural feelings: the Alpha direction, or the claim that having natural feelings entails having certain moral feelings, and the Beta direction, or the claim that having moral feelings entails having certain natural feelings. Arguing for the Alpha direction, Rawls claimed that a person who exhibited natural but not moral emotions would not be intelligible. Thus, he wrote, "it is part of the definition for a person's being proud of something, or of viewing him as holding a certain position, and considering certain things below him, that he feels shame in certain circumstances."[41] Similarly, he explained, "it is logically impossible to accept the statement that A is a friend of B, and meaning friend of B, and then to suppose that A would not feel remorse if in excessive anger etc he wrongly injured B severely."[42] Logical connections also differed from permissive or developmental explanations. Influenced by Jean Piaget whose *Moral Judgment of the Child* he assigned for the course, Rawls thought that natural emotions such as love preceded moral emotions in time and, moreover, were a "precondition for subsequent moral conduct."[43] Yet, in 1958 the laws of evolution from natural to moral emotions were in the background; his main interest was in establishing logical connections between natural and moral emotions.

Rawls's interest in these logical connections stemmed from the broader implications he could draw from them. His intention was to use the Alpha direction in order to show how incomprehensible a being lacking moral feelings would be. Having established entailment between natural and moral feelings, he made it explicit that the lack of moral feelings implies the lack of certain natural feelings: "one could not be without moral feelings without also being [without] certain natural feelings."[44] As Rawls wrote in "The Sense of Justice," and, later, in *A Theory of Justice*, imagining a person without natural feelings, we would understand that he lacks part of humanity.[45] This, he thought, would lead us to "accept our having this sense," by which he must have meant that, rare abnormal cases and young children excluded, all human beings have a sense of justice.[46] Again, hidden in these remarks was Rawls's appeal to the notion of intelligibility: a being without natural emotions or a sense of justice is not one that we understand, or regularly encounter. "If so," he concluded, "this is a kind of grounds of morality, that is, it shows what a lack of morality would involve." [47] To Rawls, this conclusion was entirely expected and only confirmed the hunch with which he started the seminar – that morality is a natural phenomenon, in the second sense of 'natural' explained above: given normal conditions, moral feelings must spring from natural feelings. This Alpha direction only made it clear that "the having of morality and moral behavior [is] rational and intelligible." [48]

[40] Rawls, "Moral Feeling 1958," Topic IX, 1i.
[41] Rawls, "Moral Feelings 1960," Seminar VI, 1i.
[42] Rawls, "Moral Feelings 1960," Seminar V, 1i.
[43] Rawls, "Moral Feeling 1958," Topic IX, 5i.
[44] Rawls, "Moral Feelings 1960," Seminar VI, 9i. The manuscript says "within," but the context requires "without."
[45] Rawls, "The Sense of Justice," in Rawls, *Collected Papers*, 112; Rawls, *A Theory of Justice*, 489.
[46] Rawls, "Sense of Justice," 112; Rawls, *A Theory of Justice*, TJ 489.
[47] Rawls, "Moral Feelings 1960," Seminar VI, 9i.
[48] Rawls, "Moral Feeling 1958," "Procedure," 2i.

The second, Beta, direction of entailment, helped Rawls established "limits [on the] content of morality" and explain the agreement of reasonable persons – should this agreement obtain.[49] Rawls reasoned about the limits to possible moral views in two complementary ways. First, he reflected on the philosophical grounds for believing that reasonable persons would agree in their judgments. We can call this a "direct" line of argument, as it focuses on the naturalist premises and then draws implications to all moral theories. Rawls was reluctant to take this line of argument because, he thought, it can soon turn into a tautological argument which defines the naturalist premises in an overly narrow way.[50] So, right away Rawls admitted that the naturalist premises could not show all but one ethical theory wrong. As he put it, naturalist premises could not be used to "settle practical moral questions in the favor of any definite code."[51] Properly understood, the naturalist premises allow for "many different types of moralities" and therefore do not by themselves solve important moral questions.[52] To solve such questions, a different type of argument was needed:

> nothing in my argument settles in advance the important moral questions of every day and politics etc in the favor of some limited and definite view. These questions, for all that I have said, are left over to be settled on their merits, and on the basis of arguments of another kind.[53]

Rawls's second, indirect, line of argument drew limits on possible moral views by focusing on individual moral theories and examining whether their particular arguments are consistent with the naturalist premises. This line of argument is familiar to readers of *A Theory of Justice*. Just as Philippa Foot had argued that the feeling of pride can be called for only in certain circumstances and therefore only certain principles can be moral principles, Rawls maintained that only certain moral principles can evoke moral (and therefore also natural) feelings. So he examined each theory one at a time, concentrating on emotivism in his seminars of 1958 and 1960. Rawls thought that the Beta argument was strong enough to exclude emotivism and existentialism as ethical theories. Arguing that one chooses moral principles and that any principle can be a moral principle, these theories were incompatible with the naturalist premises. Taking Hare as a representative of such a position, Rawls argued that, while we may need to make difficult choices, these choices are guided by beliefs that we already have: "It is really quite impossible to speak of morality as a matter of choice if what we have said is correct. There may be decisions which seem quite arbitrary within limited parts of morality – when it comes to emphasizing this value or that – but as for the idea that a person could rationally choose just this morality or that …, just choose it, independently of everything else, this idea must be wrong."[54]

Rejecting the idea that the choice of principles is entirely unguided, Rawls used the Beta direction to show just what strictures such a choice would have: the proposed moral principles would have to be connected to some of the familiar natural and moral feelings. Yet not all principles, Rawls argued, could be so connected. Principles like "do not walk on the sidewalk" could not be moral principles because they could not be connected to moral emotions, such as

[49] Rawls, "Moral Feelings 1960," Topic VII: Concept of Morality, 2ii.
[50] Rawls, "Moral Feeling 1958," Topic IX, 3i.
[51] Rawls, "Moral Feeling 1958," Topic IX, 3i.
[52] Rawls, "Moral Feeling 1958," Topic IX, 3i.
[53] Rawls, "Moral Feeling 1958," Topic IX, 3i.
[54] Rawls, "Moral Feeling 1958," Topic IX, 3i.

guilt or shame: "what I have attempted to show, after having examined some of the moral feelings, is that the standard moral feelings could not be defined with respect to any content: that is, that these feelings require certain objects."[55] Connecting these odd moral principles to natural human emotions would require a wider conceptual shift than we could accept: "the idea was to show … how very great a shift it would be in our whole way of viewing morality and human feelings etc if we assumed morality might have any content. This is [a] drastic conceptual shift."[56] Thus, in a move familiar of Wittgensteinians, Rawls claimed that a consistent emotivist was an unintelligible man.

Even though Rawls did not believe that naturalist premises were strong enough to exclude all but one ethical theories he nonetheless believed that they could explain the agreement of reasonable persons – should such agreement obtain. He believed that all moralities that are consistent with the naturalist standpoint would have some shared content. This overlap between moralities and the agreement in judgments, should they obtain, would be explained by the natural attitude of recognizing persons as persons. As Rawls wrote in an undated "Essay V," "sharing prima facie principles, there must be many types of cases on which all moralities agree."[57] In this argument, Rawls made use of Wittgenstein's notion of "family likeness" or "family resemblance": the idea that, although related practices may not share any one trait in common, they will have sufficient overlapping similarities.[58] Rawls employed the same reasoning with regard to different moral conceptions, arguing that they have a point of overlap:

> My hypothesis is this: that anything which we would call a morality has a certain specific set of prima facie principles. Or, all moralities resemble one another in their prima facie principles; they have this sort of family likeness. They resemble one another in their principles … But even though they have the same principles (or principles that bear a likeness to each other) they may differ by varying the emphasis and so favoring one principle over another, in a wider or narrower scope, and in using different frameworks….[59]

The extent of this overlap of the reasonable persons' moral conceptions was not clear, and it was still Rawls's task to discover it. However wide the overlap was, it was explained by a shared background of natural feelings:

> If it is true that moralities all have a certain set of prima facie principles in common (or some family resemblance to some set), as I think is the case, this finds its explanation in the fact that these principles are connected with forms of recognition of persons, and forms of acting with them. This set has itself between its members a family resemblance: to violate any of them would be to violate some kind of personal connection.[60]

[55] Rawls, "Moral Feeling 1958," Topic IX, 2i.
[56] Rawls, "Moral Feeling 1958," Topic IX, 2i.
[57] Rawls, "Essay V," 3, 1i.
[58] Wittgenstein, *Philosophical Investigations*, §66.
[59] Rawls, "Essay V," 3, 1i.
[60] Rawls, "Essay V," 1, 1i.

These Wittgensteinian explorations in moral feelings ended by 1964, when Rawls offered an entirely different version of the seminar on moral psychology. The emphasis of his argument shifted from logical connections between natural and moral feelings to showing how the laws of psychological development help the argument from the original position and, in particular, support the two principles of justice. Now, unlike before, Rawls's main concern was determining the relationship "between the correct psychological explanation and the correct moral theory," and asking whether all major ethical theories are compatible with correct psychological explanations.[61] Telling of the change, the rival ethical view now was not emotivism or existentialism but utilitarianism. Deciding between justice as fairness and utilitarianism required specifying "a plausible psychological theory which explains how (rational) persons acquire the desire under normal conditions to do what is right."[62] The importance of Jean Piaget, whose theories of moral development played a background role in 1958, had by 1964 significantly grown. As Rawls now understood it, moral psychology was the third, and last, task of ethical theory: while the first was to explicate considered judgments of reasonable persons and the second was to derive the principles of this explication from "philosophically defensible premises," moral psychology was meant to explain how persons can be brought to act upon these principles.[63] Accordingly, emphasis on the three psychological laws, which feature prominently in *A Theory of Justice*, had also grown. Rawls now mainly argued that, if the parents manifestly love the child, he will also develop the capacity for love and, more generally, for fellow-feeling (the first psychological law), that, if he takes part in activities the rules of which are just, he will develop the capacity for friendship and mutual trust (the second law), and, finally, if he is a beneficiary of the just and enduring practices of a society, he will develop a sense of justice, or adherence to a set of principles of justice (the third law).[64] The Wittgensteinian explorations and their conclusions remained in *A Theory of Justice*. Rawls continued to think that morality – or, by 1971, moral views – stem from "moral sentiments," where "sentiments" should be understood as "natural feelings."[65] He continued to argue that natural feelings provide the background which makes moral feelings and therefore moral reasons intelligible. And he continued to draw this connection in two directions, claiming that a person without natural feelings lacks moral feelings and consequently a person without moral feelings lacked natural feelings.[66] Nonetheless, the change from Wittgensteinian to Piagetian explorations was noticeable.

This change can be explained by citing several reasons. First, Rawls's explorations in moral psychology had a purpose other than showing one ethical theory correct. They were meant to explore the basis of morality in natural feelings. Defending justice as fairness was simply a different task. Second, the shift in emphasis can be partly explained by Rawls's positivist leanings to argue more abstractly, from the original position, and not from particular circumstances to more general conclusions. Lastly, the change must also have a more deep-seated reason: Rawls must have concluded that contextualist arguments would get us very far because they would lead to descriptions of different conceptions of various moral emotions. By

[61] John Rawls, "Moral Psychology, 1964-5," John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 6, "Introductory Remarks," 2ii.
[62] Rawls, "Moral Psychology, 1964-5," Seminar II, 4i.
[63] Rawls, "Moral Psychology, 1964-5," Seminar V, "Three tasks of ethical theory."
[64] Rawls, *Collected Papers*, 100-106.
[65] Rawls, *A Theory of Justice*, 44.
[66] Rawls, *A Theory of Justice*, 486-487.

1960, considering Hart and Hare's versions of the open-texture argument, Rawls had concluded that the same principles can be legitimately applied in different ways by different people. Naturalist premises could exclude some ethical theories, namely those that rely on emotivism and existentialism, but they could not help Rawls in his goal to show justice as fairness superior to utilitarianism. As he had argued, "many different types of moralities are allowed for under the concept of morality as we have discussed it."[67] In this regard, the argument from the original position could achieve at least as much – draw limits on possible moralities – as the argument from natural feelings: "If limits on content follow from both J as F arg[ument] and nat[ural] feelings, how are these two derivations related?"[68] Given that the original position could achieve the same and more, and that, by 1960, the argument against emotivism was clearly formulated, it is not at all surprising that Rawls started emphasizing not the logical connections between natural and moral feelings, but the laws of psychological development.

Despite this eclipse of the Wittgensteinian investigations into moral feelings by the mid-1960s, Rawls's approach to philosophy had changed. In particular, Rawls started acknowledging both the fact that reasonable persons will differ in their opinions and that, as long as these opinions are consistent with the facts of naturalism, they will overlap in significant ways. While this way of looking at agreement and disagreement allowed Rawls to claim that actual disagreement among reasonable persons is innocuous for philosophy, it affected his conception of theory and how helpful in may be in practice. This changed view of the relationship between theory and practice can be seen in Rawls's lectures on political and social philosophy.

Theory as a Guiding Framework

Rawls started lecturing on political and social philosophy at Cornell in the Spring of 1956 and continued to offer the course in every academic year until the publication of *A Theory of Justice*. The earliest notes from these lectures date to 1960; by then, they were already organized around the subjects familiar to the readers of Rawls's magnum opus: those of liberty, equality, and the common good. These notes continue many Wittgensteinian themes already seen in Rawls's seminars on moral feelings. Thus, he saw drawing conceptual, or necessary, connections between justice, other virtues of social institutions, and the concepts which justice arranges – liberty, equality, and the common good – as the main task of political philosophy.[69] These notes also contain a solidification of Rawls' early claim that the connection between theory and practice is "logically loose." If in 1954 Rawls believed that reasonable persons would nonetheless always agree on particular courses of action despite this logical looseness, by 1960 he had given up on this idea. He now argued that justice is often indeterminate and guides us at best only in a general direction. However, he continued to believe that the direction in which political philosophy guides reasonable persons is nonetheless the same.

Rawls saw political philosophy as primarily concerned with "arranging" the concepts that fell under the concept of justice: liberty, equality, and the common good. He thought of these

[67] Rawls, "Moral Feeling 1958," Topic IX, 3i.
[68] Rawls, "Moral Psychology, 1964-5," Seminar V, 3i.
[69] Rawls, "Political Philosophy 171, 1960," Lecture I, 3ii; Rawls, "Political Philosophy 171, 1962," Lecture I, 3i.

connections as logical connections, and used explicitly Wittgensteinian terms of necessity, calling the principles explaining these concepts "standards, criteria."[70] His modified description of the original position reflected this view: he started seeing it as an "analytic construction" or an "analytic framework" for "logically construing certain concepts."[71] Outlining the structure of his argument to students, he summarized the original position as a way of clarifying the conceptual connections in our reasoning: "the analytic framework which I shall use for the presentation of classical liberalism … is a rather general analysis of the concept of justice: that is, I am going to work from a certain analysis of this moral concept which is sufficiently general to allow a setting for the three notions of liberty, equality, and the common good (or, as I shall sometimes call it, social utility)."[72]

This conception of political philosophy prompted other shifts. In particular, he started seeing justice as only one virtue of social institutions.[73] Other virtues, like efficiency or humanity, expressed "distinct type[s] of moral fault" and were "associated with (or completed by) different principles."[74] For this reason, Rawls wrote, "one must avoid the tendency to identify the concept of justice with the concept of right," which, unlike justice, was associated with all the virtues of social institutions.[75] The relation of justice to other virtues of social institutions was not one of meaning, as these other virtues did not fall under the concept of justice. Rather, it was one of weight or importance: it counted for more or less than the other virtues.

This delimitation of justice brought a corresponding change in Rawls's understanding of the significance of his analysis. From his early liberal Protestant years, Rawls viewed moral philosophy as a guide to action. In 1956, he would emphasize connections between the principles of rational choice, desire and knowledge. A person who simply does not accept these principles is, according to Rawls, an unintelligible creature: "We do not understand (cannot give a sense to his saying) someone who just doesn't act on these principles; who says that he doesn't regard reasons given in accordance with them as good reasons. And the reason we cannot do so is that we think these principles involved [sic] the very concepts which we use to express ourselves in talking about choices and decisions, etc."[76] To determine a decision about any question, moral philosophy had to highlight all relevant conceptual connections, reasons to which they give rise, and suggest some weighing of these reasons. However, describing justice as only one of the virtues of social institutions, Rawls now made clear that it was also only one of the relevant considerations in the evaluation of these institutions. To render a judgment all things considered, political philosophy had to provide analyses of the remaining virtues of social institutions and determine their relative weights – a task that Rawls called analysis the concept of 'right.'[77]

---

[70] Rawls, "Political Philosophy 171, 1960," Lecture XII, 1i; Rawls, "Political Philosophy 171, 1960," Lecture II, 3i; Rawls, "Political Philosophy 171, 1962," Lecture I, 3.
[71] Rawls, "Political Philosophy 171, 1960," Lecture XII: 1i.
[72] Rawls, "Political Philosophy 171, 1960," Lecture II, 3i.
[73] Rawls, "Political Philosophy 171, 1960," Lecture III, 1i.
[74] Rawls, "Political Philosophy 171, 1960," Lecture III, 1i.
[75] Rawls, "Political Philosophy 171, 1960," Lecture III, 1i.
[76] Rawls, "Rational Choice and the Concept of Goodness," "The Grounds of Principles of Rational Choice," 6i.
[77] Rawls, "Moral Psychology, 1964-65," Seminar VIII, 1i.

In the early 1960s, the realization that analysis of justice is not sufficient to render judgment on social and political institutions increasingly troubled Rawls, in particular because he did not know how to provide the broader analysis of 'right.'[78] He did not think of appealing to the doctrine of the unity of virtues to argue that, if the social institutions are virtuous in one respect (of justice), they must also be virtuous in other respects (be humane and efficient); in fact, he often emphasized that justice was in conflict with utility, one of the virtues of social institutions.[79] To solve this problem, he decided to "try out" the idea that justice has absolute weight with respect to all other virtues of social institutions: "what I propose to do … is to try out the thought that the concept of justice does have an absolute weight, and to see whether this suggestion, in view of our considered moral opinions, leads to conclusions that we cannot accept."[80] The thought held up well; hence the opening sentences of *A Theory of Justice*: "Justice is the first virtue of social institutions, as truth is of systems of thought. A theory however elegant and economical must be rejected or revised if it is untrue; likewise laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust."[81]

The most significant change in Rawls's view of political philosophy was his new understanding of the nature of theory: he gave up the previously guiding idea that the principles of justice always help in deciding practical political questions. Elaborating on the tasks and range of questions that political philosophy covers, he wrote that practical political and social problems are "strictly speaking, outside … [its] scope."[82] Rawls meant a variety of things by this expression. He thought that the philosophical reasoning was simply more abstract and that the subject matter of political philosophy – the proper arrangement of a society's institutions and practices – simply did not consider the more particular political questions. Such particular questions required information that was not under the purview of philosophy. More importantly, Rawls also thought that, often, even if given such extraneous information, philosophers, as well as other reasonable persons, would not be able to settle practical questions conclusively. Thus he started viewing principles as indeterminate in that there were several different but equally appropriate ways of applying them in practice. As a result, philosophy did not always "make the answer calculable."[83] Often, he thought, the most that it could do was to guide political judgment in a "general direction," by providing boundaries on what can be accepted as proper interpretation of the principles; theories, he wrote, allow us to "approach [questions of justice] in a certain way, and to put certain constraints and demands on what are to be accepted as proper solutions."[84] In doing so, they often discarded many opinions as "beyond the bounds of sound opinion altogether – eg., to maximize pain; or various racist doctrines"; but even this kind of narrowing down of possible actions was not guaranteed.[85]

Rawls did not think that this conclusion was problematic: "even if the rational grounds of choice which can be found determine only a direction this is by no means to be despised; and

[78] Rawls, "Moral Psychology, 1964-65," Introductory Remarks, 2i-2ii; John Rawls, "Legal Obligation and the Duty of Fair Play" [1964] in Rawls, *Collected Papers*, 125.
[79] Rawls, *Collected Papers*, 125.
[80] Rawls, *Collected Papers*, 125-26.
[81] Rawls, *A Theory of Justice*, 3.
[82] Rawls, "Political Philosophy 171, 1960," Lecture I, 2i.
[83] Rawls, "Political Philosophy 171, 1960," "Some Notes on Use," 2ii.
[84] Rawls, "Political Philosophy 171, 1960," Lecture II, 2ii-3i, 4i.
[85] Rawls, "Political Philosophy 171, 1960," Lecture II, 2i-2ii.

may be all that is reasonable to hope for."[86] As he explained in Part II of *A Theory of Justice*, devoted entirely to application of the two principles of justice in practice, the fault was not with those who reasoned: when the decision of reasonable persons is indeterminate, we conclude that "justice is to that extent likewise indeterminate."[87] Accordingly, in cases of the indeterminacy of justice, the disagreement of reasonable persons was not rational disagreement, as reasonable persons simply did not have reasons for or against their respective interpretations of what the principles require. As Rawls wrote of different moralities, "there are many situations when judgments will differ, and we cannot say that either is correct or better, that either is wrong, given the way the person sees the situation."[88] From 1960 onward, then, he concluded that the principles of justice allow for a range of equally appropriate but differing decisions.

This change in Rawls's conception of philosophy can be explained by his appeal to Wittgenstein's notion of family resemblance. Rawls had rejected the most radical interpretation of Geach's claim that 'good' has as many meanings as its criteria: saying this, he thought, would commit us to the view that the use of the same word was "a queer linguistic fact," a view he deemed patently wrong.[89] Analysis of Rawls's examples shows that he appealed to the Wittgensteinian theme of family resemblance to argue that, despite their differences, all reasonable persons agree on at least some things. Discussing different moralities, Rawls wrote that "all moralities resemble one another in their principles; they have this sort of family likeness."[90] Yet, he continued, "even though they have the same principles (or principles that bear a likeness to each other) they may differ by varying the emphasis and so favoring one principle over another."[91] Similarly, reflecting on virtues falling under the concept of justice, he wrote that he inclined "to the view that the classical social ideals recognize the same virtues but for various reasons, moral and theoretical, assign them different interpretations and priorities."[92] Thus, appealing to the distinction between the meaning of a virtue and its weight in our overall judgment, Rawls imagined that reasonable persons all have the same conceptual framework, make the same necessary connections, only may sometimes give the same virtues different priorities.

Consistent with this picture of human reasoning is Rawls's view of disagreement. He never raised a possibility that reasonable persons may not only differ, but simply disagree, which shows how unthinkable it was for Rawls that reasonable persons might not share a broader moral conception.

Accordingly, Rawls viewed disagreement not as a fundamental disagreement between incompatible conceptual frameworks, but as a difference in rigor or clarity of expression – one that could in principle be dissipated without anyone fundamentally abandoning his or her position. His 1960 view that differences between utilitarianism and the social contract tradition are not insurmountable is a good illustration of this view. He thought that it did not matter whether one started in the utilitarian or the social contract tradition: with increased precision one

[86] Rawls, "Political Philosophy 171, 1960," Lecture II, 2ii-3i.
[87] Rawls, *A Theory of Justice*, 200.
[88] Rawls, "Essay V," 3: 1.
[89] Rawls, "Rational Choice and the Concept of Goodness," 9: 6ii.
[90] Rawls, "Essay V," 3: 1.
[91] Rawls, "Essay V," 3: 1.
[92] Rawls, "Political Philosophy 171, 1962," Lecture I, 3.

would nonetheless end up with the same conclusions. This view that the history of philosophy is a history of cumulative development was typical of analytic philosopher of the period. Like them, Rawls thought that progress was market by increased precision. "I look at the development of political philosophy," Rawls wrote, "as the development of more precise understanding of moral concepts and principles as they apply to political questions."[93] Historical figures, he thought, worked with the same or similar concepts; thus, the utility of reading Aquinas was in seeing how previous accounts of the concept of justice were "incomplete."[94] Aquinas's account of justice was "not so much incorrect, but … not as strong as one would like: that is, it fails to provide a complete account of our judgments about justice."[95] In particular, Aquinas fell short in thinking that all his moral injunctions followed from the concept of natural law, and in being vague in his conception of the common good."[96]

Accordingly, a contemporary political philosopher had the task of improving the rigor and breadth of previous thinking: "What one should try to do is to make a substantial improvement over what has gone before. … it will be a start simply to collect together and try to answer in a consistent way the main questions to be answered in giving an analysis of justice."[97] To do so, she had to ask how to build on Aquinas's achievements and search for ways to "strengthen and improve Aquinas's account."[98] Indeed, Rawls saw himself as building on previous achievements; as he wrote in 1960, he approached political philosophy "by asking what are the least changes and amendments which have to be made in the utilitarian tradition, or in some utilitarian writer … to render the view stated true…."[99] Taking utilitarianism as the starting point was not significant for Rawls; as he pointed out immediately afterwards, one could get to the same conclusion by starting with the social contract tradition: "one could equally profitably, I suspect, begin with Rousseau or Kant asking the same question [the least changes one could make] and end up, in either case, in much the same place."[100] In sum, despite introducing difference in the judgments of reasonable persons, Rawls never thought that they diverged in their conceptual frameworks.

Some passages from the period suggest that Rawls was not always comfortable with this conclusion. He obviously wanted to say that utilitarianism differs from his two principles in more fundamental ways than placing different weights on the same virtues. Explaining how one may recognize the same virtues and yet act differently, he gave an example of different economic policies. While "most everyone would agree on the desirability of efficient allocation of resources (per period), full employment, high rate of growth of GNP, price stability, all consistent with distributive justice and liberty," attaching different weights to these goals, different people would recommend different economic policies.[101] While the example was perfectly consistent with his explanation, Rawls added that these different economic policies

[93] Rawls, "Political Philosophy 171, 1962," Lecture I, 1i.
[94] Rawls, "Political Philosophy 171, 1962," Lecture VIII, 1i.
[95] Rawls, "Political Philosophy 171, 1962," Lecture VIII, 1i.
[96] Rawls, "Political Philosophy 171, 1962," Lecture VIII, 7i.
[97] Rawls, "Political Philosophy 171, 1962," Lecture I, 1i.
[98] Rawls, "Political Philosophy 171, 1962," Lecture VIII, 1ii.
[99] Rawls, "Political Philosophy 171, 1960," Lecture II, 5ii.
[100] Rawls, "Political Philosophy 171, 1960," Lecture II, 5ii.
[101] Rawls, "Political Philosophy 171, 1962," Lecture I: 3.

"presumably express different (underlying) social ideals."[102] This remark shows his hesitation in maintaining that the conceptual frameworks of reasonable persons are fundamentally the same.

Conclusion

Rawls's dilemmas from the period were set by debates about the nature of necessity. In the seminars on moral feelings, Rawls elaborated the nature of necessary connections as conceptual connections without which the use of a word was unintelligible. Relying on this conception of necessity, Rawls argued against the emotivists, showing that moral feelings are necessarily connected to certain moral reasons and natural feelings, thereby concluding that not any principle can be a moral principle. In lectures on political philosophy, this notion of necessity set Rawls's concerns, and, throughout the course, he drew necessary connections between justice and the attendant concepts of liberty, equality, and the common good. He elaborated the notion of sameness which allowed for differences in criteria; here as well, he relied on the Wittgesnteinian theme of family resemblance, claiming that all moralities could be called by the same name as long as they contained some of the family of moral principles. He allowed the same for his own principles of justice: reasonable men could be said to accept his principles of justice as long as their application of these principles bore a certain family likeness. The result of this modification was the changed role of political philosophy in practical political affairs: it offered only a general direction to action and set only general constraints on the application of principles.

By the end of this period, the core commitments of Rawls's theory had already been set, yet one important question remained. While he maintained the universality of his theory and claimed that all reasonable persons would agree in their judgments, he needed to specify ways in which theories may fail to explain such judgments if all reasonable persons do in fact fundamentally agree. A failure to explain these judgments would be in many ways unexpected; Rawls himself, upon asking what he would do if his analysis of our considered judgments was wrong, wrote that this "would be rather odd," because it would mean that "[his] considered judgments were different from most, or nearly all, other competent persons."[103] For the same reason, however, the failure of utilitarians to analyze our moral experiene should also be odd. By 1965, when the influence of Rawls's colleague at Harvard, W.V.O. Quine, began to be felt, Rawls would deal with this question by making use of Quine's explication as elimination.

[102] Rawls, "Political Philosophy 171, 1962," Lecture I, 3.
[103] John Rawls, "Moral Judgment, Relativism" [1958], John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 8, 7: 1i.

6

Re-emergence of Positivism

The period between 1962 and 1971 contains many developments in Rawls's thought. If by the early 1960s Rawls still viewed ethics from a recognizably positivist perspective, treating the subject as an empirical inquiry aimed at explicating considered judgments of reasonable persons, in 1971 he would describe his argument as Kantian and part of the rational choice and social contract traditions. It is therefore a real question whether these developments do not mark a change in Rawls's conception of philosophy, and whether *A Theory of Justice* is not better explained with reference to the traditions he then emphasized.

In this chapter, I want to survey Rawls's intellectual development from the early 1960s to 1971 keeping this question in mind. I argue that, despite the influence of said traditions, Rawls's argument remained positivist. As I show in the first two sections, Rawls's positivism was in fact reinforced by his Harvard colleague W.V.O. Quine, "the greatest logical positivist."[1] By 1965, Rawls's conception of philosophy became expressly non-foundational, but, unlike contemporary historicist approaches, it avoided the flux of the non-foundational world by positing fixed points – points of the agreement among reasonable persons. This expected agreement informed Rawls's account of justification: he saw it as a matter of consistency – reflective equilibrium – between these fixed points and the principles of justice.

In the next three parts of the chapter, I explain how Rawls attempted to uncover this state of reflective equilibrium by way of a thought experiment between 1952 and 1971. I provide the history of this thought experiment in order to show the positivist aim that persists in it despite Rawls's later re-description of it as part of the Kantian, social contract, or rational choice traditions. This positivist aim – to explicate the judgments of reasonable persons and to uncover agreement therein – illuminates ways in which Rawls's argument did indeed become Kantian and used tropes from the social contract and rational choice theories while remaining positivist in its conception of philosophy.

Quine, the positivist

Rawls first joined Harvard as a visiting professor in 1959-60, and, after spending two years at the Massachusetts Institute of Technology, returned there as a full professor in 1962. Although by the early 1960s Harvard had become a center for Wittgensteinian thought, Rawls's views were most markedly influenced by Quine.[2] Rawls met Quine during his first stay at

[1] Putnam, *Realism with a Human Face*, 268.
[2] From 1962 onward, Harvard employed David Sachs, Rogers Albritton, Burton Dreben, and Stanley Cavell, all of whom were influenced by Wittgenstein's thought.

Harvard, read Quine's books, most of them in entirety, and discussed them with Harvard colleague Burton Dreben, who on Rawls's own account made "Quine's view clear" to him and with whom he worked intensely between 1962 and 1967.[3] Although Quine wrote virtually nothing on normative subjects and seems to have had little personal interaction with Rawls, he brought Rawls closer to the positivist position with which he started. This is not a coincidence: Quine belonged to the logical positivist tradition both in his intellectual origins and in the character of his later commitments. As a young philosopher, he studied with positivists Rudolf Carnap in Vienna and Alfred Tarski in Warsaw in the academic year 1932-33. He never fully agreed with Carnap, but he nonetheless saw Carnap as "the leader of the continuing developments" in philosophy from the 1930s onward.[4] Despite their disagreement, Quine thought that Carnap "was still setting the theme," and that his own "line of … thought was largely determined by problems that … [Carnap's] position presented."[5]

Many of Quine's core commitments thus stemmed from the positivist tradition. Most broadly, Quine's approach to knowledge was empiricist in its reliance on data acquired by the senses: he held that "physical things generally, however remote, become known to us only through the effects which they help to induce at our sensory faculties."[6] Like Carnap's later position, Quine's empiricism was anti-foundational: he did not believe that knowledge gained by sensory qualities is unquestionable or necessary.[7] As his criticism of foundationalism in "Two Dogmas of Empiricism" [1951] reveals, Quine's arguments were anti-foundational because of meaning holism, or the claim that the meaning of any one term depends on the meaning of other terms. Quine offered two arguments against foundationalism in "Two Dogmas": that the notion of analytic and necessary truths is not clearly defined, and that foundationalism's key premise – the reduction of all knowledge to immediate and defined experiences – is flawed because knowledge is not stored in individual statements or experiences but rather "the unit of empirical significance is the whole of science."[8] In short, Quine argued that foundationalism's key premise was shown wrong by meaning holism.

The key implication of meaning holism was justificatory holism, or the claim that one justifies not any single statement of a theory, but the theory as a whole. Any one statement – including the allegedly necessary statements – does not have many implications by itself: "a scientific sentence cannot in general be expected to imply empirical consequences by itself. A

---

[3] Rawls, 'Autobiographical Notes', 21. Rawls, *A Theory of Justice*, xi. For Rawls's annotations of Quine's books, including *From a Logical Point of View* [1953], which Rawls read in entirety, *Word and Object* [1960] (read in entirety), *The Ways of Paradox and Other Essays* [1966], *Ontological Relativity and Other Essays* [1969]. For these annotated books, see John Rawls Personal Library, Harvard University Archives, HUM 48.1, Box 6.
On Rawls's reflections about Burt Dreben and Dreben's influence on him, see John Rawls "Afterword: A Reminiscence" in Juliet Floyd and Sanford Shieh, eds. *Future Pasts: The Analytic Tradition in Twentieth Century Philosophy* (Oxford: Oxford University Press, 2001), 417-430, especially p. 423, where Rawls states, "I can't think of any of my basic ideas that I got from Burt, yet I am convinced that replying to his criticisms always enormously improved the clarity and the organization of my thought." For Rawls's collaboration with Dreben, see Ibid., 424.
[4] W.V.O. Quine, "Homage to Carnap" in *Dear Carnap, Dear Van*, ed. Richard Creath (Berkeley, CA: University of California Press, 1990), 463-4.
[5] Quine, "Homage to Carnap," 463-4.
[6] W.V.O. Quine, *Word and Object* (Cambridge, MA: the MIT Press, 1960), 1.
[7] See, for instance, Quine, *Word and Object*, 22.
[8] W.V.O. Quine, "Two Dogmas of Empiricism," *The Philosophical Review* 60 (1951): 39.

bigger cluster [of assumptions] is usually needed."[9] As a result, by testing any one statement, we are in fact testing the "bigger cluster" of premises on which the statement relies. In short, Quine argued that theories stand the test of experience not as a collection of individual statements, but as a collection of interdependent premises.[10]

Quine's version of meaning holism was radical in its implications for positivism's analytic-synthetic distinction, but it remained indebted to the tradition's key commitments. In particular, Quine continued to believe that meaning holism would not harm positivism's claim that all scientific observers would agree on at least some scientific statements. He followed the tradition in calling these statements "observational statements," or statements to which other beliefs are largely irrelevant. As he put it, observational statements are statements "most strongly conditioned to concurrent sensory stimulation" and least dependent on our wider web of beliefs, or "stored collateral information" or "stored information beyond what goes into understanding the sentence."[11] Being least dependent on the wider web of beliefs, observational statements were also those "on which all speakers of the language give the same verdict when given the same concurrent stimulation."[12] In effect, then, although Quine endorsed meaning holism, he limited its implications by allowing that some observations are little affected by the wider webs of beliefs of those who observe. This feature of Quine's thought is little emphasized: noted for his meaning holism, Quine is thought to have opened the door to contesting the existence of observational sentences. In fact, however, Quine did not take that step. In that respect, he remained a firm positivist.

The extent of Quine's positivism is most apparent in contrast to the contemporary historicist approach to the philosophy of science. This latter approach, best exemplified by Thomas Kuhn, drew significantly more radical implications from meaning holism. Like Quine, Kuhn criticized the early logical positivist understanding of observation by arguing that individual observations took place in the context of a wider scientific theory.[13] Unlike Quine, however, Kuhn objected to the notion of "observational statements" defined in terms of sensory impressions, claiming that questions about sensory impressions "presuppose a world already perceptually and conceptually subdivided in a certain way."[14] Starting from these different premises, he criticized Quine for assuming that "two men receiving the same stimulus must have the same sensation."[15] In the absence of observational statements or other shared beliefs, scientific theories could not be judged to be better or worse by appealing to these commonly shared beliefs. Instead, Kuhn wrote, scientific theories are justified from their own point of view

[9] W.V. Quine, "Two Dogmas in Retrospect," *Canadian Journal of Philosophy* 21 (1991): 272.
[10] W.V. Quine, "Epistemology Naturalized," in W.V. Quine, *Ontological Relativity and Other Essays* (New York: Columbia Press, 1969), 79.
[11] Quine, "Epistemology Naturalized," 85-6.
[12] Quine, "Epistemology Naturalized," 86-7.
[13] Kuhn, *Structure*, 125-29.
[14] Kuhn, *Structure*, 129.
[15] Kuhn, *Structure*, 202f.

and with reference to their fruitfulness in explaining the world.[16] Similar positions were held by Michael Polanyi and Norwood Russell Hanson.[17]

Quine dismissed this radical interpretation of meaning holism and defended the notion of observational sentences. Claiming that thinkers such as Kuhn, Polanyi and Hanson "belittle the role of evidence and … accentuate cultural relativism,"[18] he argued that one could have observational statements that encompass the entire scientific community: "what counts as an observation sentence varies with the width of community considered. But we can also always get an absolute standard by taking in all speakers of the language, or most."[19] He seemed certain that this universal standard would be met. As the contrast to historicism shows, Quine's "observation sentences" were meant to play the role of "hard" evidence: evidence that, while not foundational, was little dependent on wider frameworks of beliefs and therefore provided common points to adjudicate between them.

Rawls drew on this positivist conception of scientific inquiry in his ethical arguments, but he did so against Quine's own judgment. Quine thought that meaning holism was so pervasive in ethics that this discipline contained no observational statements and therefore no subject matter. Ethical statements such as "that's outrageous" are not observational statements because their truth or falsity "hinges on collateral information not in general shared by all witnesses of the acts [that are said to be outrageous]."[20] Paradoxically perhaps, despite rejecting the thought of modeling ethics after scientific inquiry, Quine did not espouse ethical relativism. Rather, he thought all societies would agree on at least some ethical principles, on some "common core," because "the most basic problems of societies are bound to run to type."[21] In ethics, just as in his entire approach to philosophy, Quine did not think that competing intellectual traditions would present incommensurable webs of belief.

Rawls's Anti-Foundational Themes: Justification and Reflective Equilibrium

Quine's influence started showing in Rawls's work in the early 1960s. Rawls accepted meaning holism in Quine's limited sense, although unlike Quine he concentrated not on the epistemological arguments but on the implications meaning holism had in ethics. In particular, Rawls emphasized justificatory holism and argued against Cartesianism that a single consideration is insufficient to deduce principles of justice. Rawls's acceptance of meaning

[16] Kuhn, *Structure*, 198-210.
[17] Michael Polanyi, *Science, Faith, and Society* (London: Geoffrey Cumberlege, 1946); Norwood Russell Hanson, *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science* (Cambridge: Cambridge University Press, 1958).
[18] Quine, "Epistemology Naturalised," 87.
[19] Qune, "Epistemology Naturalised," 88.
[20] W.V.O. Quine, "Reply to Morton White," in Lewis Hahn and Paul A. Shilpp, eds., *The Philosophy of W.V. Quine* (La Salle, IL: Open Court, 1986), 664.
[21] W.V.O. Quine, "On the Nature of Moral Values," In W.V.O. Quine, *Theories and Things* (Cambridge, MA: Harvard University Press), 62. It did, however, mean that "a coherence theory of truth is evidently the lot of ethics." Quine, "Moral Values," 63.

holism started showing in his extended criticisms of foundationalism: a topic that had not appeared in Rawls's writings before this period. His understanding of foundationalism was typical in the contemporary context: he saw foundationalism as "Cartesianism," or an attempt to claim certain premises as self-evident and necessary, and then deduce ethical conclusions from these premises alone. As he explained to his students in 1966,

> There is a tradition in philosophy – let's call it Cartesianism – which thinks of justifying a proposition as deducing it from self-evident premises, from necessarily true statements. Taking statements of concept identity, logic and mathematics as such statements, we might try to justify our ethical [conclusions] from these.[22]

Rawls rejected foundationalism for two main reasons. First, he thought that no adequate account of necessity had yet been given. As he argued in the 1967 lectures on ethics, any account of necessity has to be placed in a larger, philosophical framework which makes clear the implications of thinking something necessary:

> In general I agree with Quine, or at least as I understand him, that no one has yet given a philosophically useful account of logical or mathematical necessity which distinguishes it and shows why it [is] essential philosophically to show that certain propositions are necessary in this sense. No doubt we can take as given by enumeration a class of (logical) truths and definitions and then clarify this class and its consequences as logically or mathematically necessary. But why this class, especially with its definitions, is of any particular significance has yet to be explained.[23]

As Rawls thought that no philosophically useful account of necessity had been given, he saw no reason to accept or deny attempts to declare certain qualities as part of a definition of 'justice,' 'good,' or any other concept. This can be best seen in his remarks in a 1965 seminar on the good. Arguing against the students who held that being moved by something judged good is part of the definition of 'good,' Rawls claimed that there is no significance – and therefore no reason – to include something as part of a definition:

> You want to make it part of the concept of goodness that to recognize that X is good implies being moved to some degree. I believe that in the absence of an account of necessary truths which shows why this connection is desirable from a theoretical point of view, little if anything is gained. And I don't believe that we have an adequate account of necessary truths.[24]

Rawls allowed for a theoretical possibility that such a philosophically useful account of necessity would be elaborated in the future, but he could not imagine even a rough structure of such an

[22] John Rawls, "Political Philosophy 171, Lectures I-IV 1966-1967." (1966) John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 36, Folder 10, "A question of Justification."
[23] John Rawls, "Analytic Ethics and Its Justification, 1966-1967" John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 5, Folder 6, 1967 Ethics, 2ii.
[24] John Rawls, "Goodness as Rationality" [1965]. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 17, 1i-1ii.

account. This seeming impossibility stemmed from the fact that there are always several feasible definitions to any concept used in the theory, and the choice of any one of these definitions needs reasons. As Rawls wrote in 1967:

> if we specify correct moral principles as those which would be agreed to by rational men, we need a (real) definition presumably of the concept of a rational man. But as there [are] various interpretations of rationality (as well as of other notions we would have to rely on), we might just as well take our preferred interpretation as an extra premiss, and drop the pretense that our conclusion is in any way necessary. … We should abandon, at least in ethics, the idea that philosophy is the analysis of concepts.[25]

Rawls deemed this first argument against foundationalism sufficiently strong to drop the notion of necessity from his philosophical repertoire.[26] Yet he did not leave the argument against foundationalism at that, and further argued that foundationalism is wrongly reductionist in its attempt to derive ethical conclusions from a small number of purportedly necessary premises. This argument stemmed from meaning holism, which was implicit in Rawls's thinking since his positivist years. Now Rawls made his meaning holism explicit. As Quine argued that "a scientific sentence cannot in general be expected to imply empirical consequences by itself," so Rawls claimed that the allegedly necessary ethical premises are not sufficient by themselves to yield a conception of justice:[27]

> there is no hope [to derive ethical conclusions] without complex definitions which [are] in effect further premises and not in any way necessary. (Quine on unclarity of analytic and the notion of concept identity.) There may be value in the Cartesian exercise, but it doesn't provide a Cartesian justification based upon necessary truths alone.[28]

Again:

> In philosophy we are all too prone to jump straightaway to the problem of justification. Now if we could produce a deduction of a complete set of moral principles (including principles of justice) from self-evident (or clearly true) premises, there would be no problem. But I am doubtful that such a Cartesian justification is … possible; digging down to basic assumptions is unlikely to give the requisite self-evident premises.[29]

Rawls deemed such deduction from self-evident and necessary premises was unlikely mainly because arguments that proceed from self-evident principles and truths of logic do not say anything about human life:

---

[25] Rawls, "Analytic Ethics," Ethics 169 (1967), 2ii-3i.
[26] In *A Theory of Justice*, Rawls explicitly stated that the principles of justice are "contingent, in the sense that they are chosen in the original position in the light of general facts." Rawls, *A Theory of Justice*, 578.
[27] Quine, "Two Dogmas in Retrospect," 272.
[28] Rawls, "Analytic Ethics," "1966 Philosophy 191," "A question of Justification."
[29] Rawls, "Political Philosophy 171, 1966-67," Lecture I, 3ii.

> The truths of logic are truths about very general notions: propositions, individuals, properties, relations; and about certain (logical) relations given by enumeration. It is not likely that truths of this kind about such general notions suffice to determine what our ethical principles should be, what a rational man should accept.[30]
>
> …
>
> One doesn't want a justification rooted in logic alone. That would only show that morals had nothing to do with men…."[31]

Rawls played a devil's advocate to ethical Cartesianism by suggesting that it take as its premises the human purposes "which would be self-contradictory not to have."[32] That would be a step in the right direction, he agreed, but added that, even "if there are such purposes, they will not suffice to vindicate and give content to a system of ethical principles."[33] Already by the early 1960s, Rawls was convinced that, devoid of claims about human life, Cartesianism had little to contribute to discussions about justice. At most, the self-evident claims would be part of a broader ethical argument.

Rejecting foundationalism, Rawls detailed his own non-foundationalist approach to ethical questions. This approach shows the unmistakable influence of Quine, and, this way, the continued guiding influence of positivism in Rawls's thought. Rawls's non-foundational approach consisted of three key claims: a conception of justice is justified if it is supported by many kinds of considerations; it is therefore justified as a whole or, as Quine put it, a theory stands the test of experience as a whole, by showing how all of its parts are supported by these many kinds of considerations; and, finally, a conception of justice is justified not absolutely but relative to other conceptions of justice. Most generally, a conception of justice is justified not by appealing to necessary truths, but by showing that its principles are consistent with all the provisional fixed points accepted by reasonable persons: a state which Rawls called "reflective equilibrium."

The most emphasized part of Rawls's non-foundational approach was his belief that theories of justice are evaluated by many kinds of considerations. As he wrote in the 1965 draft of *A Theory of Justice*, "the justification of a conception of justice is almost certain to be cumulative and to rest on the consilience of many distinct considerations."[34] The emphasis on the scope of considerations required for ethical arguments was not entirely new to Rawls: he argued against Kurt Baier and Richard Hare that formal conditions on the concept of justice – universalizability and prescriptability – are not sufficient by themselves to deduce a conception of justice. However, if in the late 1950s Rawls concentrated on showing that these formal conditions are inadequate, in the 1960s he made this argument with respect to all kinds of considerations.[35] The claim that in moral philosophy "there are no shortcuts of this sort" – no appeals to special kinds of considerations – became pervasive in Rawls's approach to questions

[30] John Rawls, "Philosophy 169. Lectures I-IV." John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 5, Folder 2. Lecture IV, 6i.
[31] Rawls, "Analytic Ethics," "1966 Philosophy 169," 9i-9ii.
[32] Rawls, "Analytic Ethics," "1966 Philosophy 169," 8ii-9i.
[33] Rawls, "Analytic Ethics," "1966 Philosophy 169," 8ii-9i.
[34] John Rawls, "Philosophy 171. Chapters on Justice. Draft of *A Theory of Justice* reproduced to students" (1965 Fall). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 18, Folder 4, 2-3.
[35] Rawls, "Analytic Ethics," "1966 Philosophy 169," 8ii-9i.

of justice. Rawls stated that considered judgments are not sufficient by themselves for the philosopher in ethics: "we must assume that the fixed points [considered judgments] are not sufficient to eliminate all but one set of principles. Several alternatives will presumably remain."[36] Other types of considerations were not sufficient for this purpose either. These included, as we have seen in this and the earlier chapter, truths of logic and definition, the formal conditions imposed on the concept of justice, and truths of moral psychology. All of these, taken singly, provided "too slender a basis" for arguments of justice.[37]

The second commitment of Rawls's non-foundational approach to philosophy was that conceptions of justice are justified as wholes. This requirement was for the most part a presupposition of the first one: many kinds of considerations were required to support a conception of justice because conceptions of justice had many ingredients which depended on different kinds of knowledge. To propose a feasible conception of justice one had to combine many kinds of considerations. But secondly, Rawls also believed that disputes on any particular topic cannot be resolved without reference to the implications of the disputed notions in other areas of thought. He drew an important implication from this requirement of holistic justification: since conceptions of justice depended on many considerations, any conception of justice was bound to be contested and possibly found wrong somewhere. Consequently, Rawls insisted, one could not fault a theory merely for being wrong somewhere. As he wrote in *A Theory of Justice*, "objections by way of counterexamples are to be made with care, since these may tell us only what we know already, namely that our theory is wrong somewhere. The important thing is to find out how often and how far it is wrong."[38]

This second commitment and its implication that a conception of justice is bound to be weak or wrong somewhere led Rawls to the third, perhaps the most important truth of non-foundational justification: conceptions of justice are justified not absolutely, by showing that they are entirely right in the lights of our current beliefs, but relatively, by showing that one conception of justice is better than others because it has fewer weaknesses than these rival conceptions. This truth of non-foundational justification is first mentioned in Rawls's 1964 seminar on moral psychology, where, referring to William Frankena's "Obligation and Motivation," Rawls wrote:

> At the end of his essay ["Obligation and Motivation in Recent Moral Philosophy"] [Frankena] suggests that the dispute between externalism and internalism in ethical theory cannot be resolved by small-scale investigations taking into account only a fragment of the problems involved. He thinks that 'each theory has strengths and weaknesses, and deciding between them involves determining their relative total values as accounts of morality. But such a determination calls for a very broad inquiry.' … I should like to second this

[36] Rawls, "Analytic Ethics," "1966 Philosophy 169," 11ii.
[37] John Rawls, "Philosophy 169. Part II. Lectures V-IX" (1970). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 5, Folder 3, Lecture VIII, 3b. For Rawls's reasons to call these conditions 'formal,' see Ibid., Lecture VII, 2ii-3i: "The term 'formal' is merely a label. That is, I shall not attempt to define it, and I do not know how to do so. Presumably, since it seems a natural term to use, it expresses some intuitive notion; but what this notion is I am not able to say. Thus by formal conditions I mean those conditions on first principles that are given by a certain list (by enumeration)."
[38] Rawls, *A Theory of Justice*, 52.

> opinion: it is often possible to decide between views if one broadens the lines of investigation; or more likely perhaps each view will turn out to be inadequate in some way.[39]

Rawls would state explicitly his belief that justification in ethics is relative in his later writings. As he told his "Political and Social Philosophy" class in 1965,

> Philosophy proceeds by argument against other positions in large part. In this sense it is dialectical. We see the weaknesses and strengths of our own position by comparing it with other positions. Thus our aim is to ascertain where A's position is weak so that we may try to go beyond it in these respects.[40]

This third truth of justification led Rawls to a realization that would have significant effects for Anglo-American philosophy. Since justification was relative, Rawls thought, it must involve comparison of rival theories; this comparison, in turn, requires the elaboration of rival theories so that their relative strengths and weaknesses are seen. As Rawls told his "Ethics" students in 1967,

> the justification of an ethical conception rests on the consilience of many kinds of considerations and upon a judgment of the relative advantage of one set of ethical principles over another. We are not in a position to judge between ethical conceptions (that is, systems of moral principles) until we know a great deal about the substantive structure of particular views – and much [more] than we know now.[41]

This truth of non-foundational justification led to a fresh way of dealing with questions of justice and ethics more generally: it started with moral questions and then compared rival answers to these questions. Against the background of contemporary moral philosophy which commentators called dead, moribund or boring in an "original way" since it did not discuss moral questions at all, Rawls's approach to questions of justice was novel and fresh.[42] Indeed, Peter Laslett, having proclaimed moral philosophy dead in the 1950s, included Rawls's "Justice as Fairness" as an example of philosophy that was reviving the discipline.[43] Rawls himself did not share Laslett's views, nor did he think that moral philosophy was dead. But, like Bernard Williams who criticized contemporary moral philosophy for being boring, he chided the contemporary trend to start ethical studies with the issues of justification. "This emphasis on justification, at least at the outset, is unfortunate" he wrote in 1967, precisely because "we are not in a position to judge between ethical conceptions … until we know a great deal about the substantive structure of

---

[39] Rawls, "Moral Psychology, 1964-1965," Seminar I, 2i. The essay referred to is William K. Frankena, "Obligation and Motivation in Recent Moral Philosophy" in A.I. Melden, ed. *Essays in Moral Philosophy* (Seattle: University of Washington Press, 1958), 40-81.

[40] John Rawls, "Natural Law, 1962, 1965" (1965). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 13, Lecture VIII, 1i-1ii.

[41] Rawls, "Analytic Ethics," "1967 Ethics 169," Lecture of Justification, 1i.

[42] According to Williams, "Contemporary moral philosophy has found an original way of being boring, which is by not discussing moral issues at all." Bernard Williams, *Morality: An Introduction to Ethics* (Cambridge: Cambridge University Press, 2011 [1972]), xvii.

[43] Laslett and Runciman, *Philosophy, Politics, and Society, 2nd Series*.

particular views."[44] Rawls's approach to political and ethical questions was one of the few fresh breaths of air from that perspective.

Ironically, however, Rawls's emphasis on substantive ethical questions showed the influence of the very tradition that made moral philosophy dead according to Laslett: positivism. Rawls's indebtedness to this tradition is shown best in his continuing – albeit now more explicit – reliance on the shared space of agreement between reasonable persons to rectify some of the instability of the non-foundational world. Rawls introduced the notion of "fixed" points in our judgments. Although a new notion, it was merely a name for an idea long implicit in his thought. Unlike judgments that are only considered (made upon due reflection, in situations where one's own interests are not involved), the "fixed points" are judgments that are also shared by all rival ethical theories. They are "certain obvious or common sense judgments which we suppose are true (or correct) beyond question. … [they] are assumed to be accepted by all; they are not in doubt."[45] Like the logical positivists and especially Quine, Rawls relied on non-foundational fixed points that played the role of "hard evidence."

In sum, justification of the principles of justice consisted in showing that, taken together, these principles are consistent with the provisional fixed points and other truths accepted by reasonable persons – or, at least, more consistent than rival principles of justice. In 1962, Rawls introduced the phrases "equilibrium of reflection" and "reflective equilibrium" to define this state of consistency.[46] Again, this was a new name for an old idea – dating at least to Rawls's years at Oxford – that, in a world without necessary truths or objectively existing ethical concepts, consistency was the key criterion for justification of ethical views. "What one is trying to achieve," Rawls wrote in his lectures on political philosophy, "is a state of <u>self-conscious reflective equilibrium</u> with respect to one's own judgments on the justice and injustice of institutions (acts, and persons)."[47]

By the mid-1960s, as Rawls made it explicit that justification of a conception of justice rests on a consilience of many kinds of considerations, the state of equilibrium was expanded to include not only the fixed points in our considered judgments but also other matters relevant to questions of justice, including the truths of moral psychology and the implications of having a morality. Rawls had settled his opinion on these different pieces of the equilibrium by the early 1960s: he had stated the principles of justice in "Justice as Fairness" [1958], had settled his views on the laws of moral psychology by 1962-64 and on the implications of having a morality by the late 1950s. Now he needed an argument to connect these pieces and see what principles of justice they imply. Started in the mid-1950s and eventually known as the "original position," this argument would become Rawls's main focus from 1964 onwards.[48]

---

[44] Rawls, "Analytic Ethics," "1967 Ethics 169," Lecture of Justification, 1.

[45] Rawls, "Analytic Ethics," "Philosophy 169," Lecture IV. I am not entirely certain of the location of this document, partly because Rawls's papers have been reorganized since I first visited them. An alternative location may be in Rawls, "Philosophy 169, 1970 Fall," Lecture IV.

[46] For "equilibrium of reflection," see John Rawls, "Political Philosophy 171, 1962," Lecture XVI, 2i. For "reflective equilibrium," see Rawls, "Draft of *A Theory of Justice* reproduced to students, 1965," 18.

[47] Rawls, "Political Philosophy 171, 1962," Lecture IX, 2i.

[48] Rawls, "Political Philosophy 171, 1962," Lecture XVII, 3ii. This term first appeared in print in "Constitutional Liberty and the Concept of Justice" (1963). See Rawls, *Collected Papers*, 73-95.

The "Original Position"

The original position was a thought experiment, or, as Rawls called it in the early 1960s, an "analytic construction," meant to collect the considerations relevant to questions of justice and determine what principles of justice, combined together, these considerations imply. As Rawls put it in *A Theory of Justice*, the argument from the original position aimed to "collect together into one conception a number of conditions on principles that we are ready upon due consideration to recognize as reasonable" and then "establish that taken together they [impose] significant bounds on acceptable principles of justice."[49]

The idea to create a thought experiment in order to explicate considered judgments of justice dates back to the period between 1952 and 1954. There are four significantly different versions of this experiment: the first elaborated in the 1952-4 notes, the second in the 1958 article "Justice and Fairness," the third in the 1964 and 1965 drafts of *A Theory of Justice*, and the final version in *A Theory of Justice* in 1971. In all these versions, the goal of the thought experiment was to reveal the principles underlying considered judgments of justice. In that respect, the thought experiment in all of its versions should be considered as part of the positivist tradition which regards considered judgments as the subject matter of the theory and which treats the overlap of these judgments as an indication that these judgments are objective.

However, the history of the thought experiment shows that Rawls's positivism became less and less apparent over the years. The phrase "subject matter of ethics" disappeared from his philosophical vocabulary, while his theory's universalism and the notion of objectivity received new interpretations that were foreign to positivism. The four versions of the thought experiment mark different steps in the retreat of Rawls's positivist self-description, and they associate the thought experiment with different traditions of thought. The first three versions of the thought experiment show the influence of the theory of games, the social contract tradition and Kantianism, respectively. The last, fourth, version contains all these influences; it does not differ from its 1965 predecessor in any significant way, but, being the final and the most developed version of the thought experiment, it deserves a section of its own.

Despite these new self-descriptions, Rawls's theory preserved its positivist features over the years. The new Kantian and social contract associations which Rawls gave to the positivist requirement of universality and its notion of objectivity did not actually make his theory Kantian or contractualist – even though they added Kantian and social contract elements to his thought. The first version of the thought experiment dates back to Rawls's analysis of considered judgments of justice in 1952, when he composed the "pure case" experiment. Rawls's goal was to reveal principles implicit in the considered judgments of reasonable persons. To do so, he modeled the thought experiment to reflect the considered judgments of reasonable persons and situations in which such judgments are made.[50] Much more explicitly than in the later years, in 1952 Rawls followed the positivist trajectory: if the thought experiment is successful, he thought, it would show that all reasonable persons agree in their considered judgments. This, in turn, would be a proof that ethical judgments were objective.

---

[49] Rawls, *A Theory of Justice*, 21, 18.

[50] See Chapter Four of this Dissertation, section "Rawls at Oxford and Cornell."

Although the idea behind Rawls's thought experiment was positivist, it contained elements of game theory already in its inception. In particular, Rawls was impressed by game theory's conclusiveness: games, he wrote, are usually "decidable": in typical circumstances, given the rules of the game, a winner can be determined.[51] As Rawls attributed this conclusiveness to game theory's use of simplifying devices such as a chooser whose rationality was clearly defined, he decided to use these simplifying devices for his own purposes. But, as we will see in the later discussion, the key premise of Rawls's experiment was not the definition of rationality but rather that of 'reasonableness' – which was absent from game theory works. Thus, although in his 1953 Oxford notes, Rawls's thought experiment featured rational egoists tasked with choosing principles for social institutions, these rational egoists were compelled to be reasonable in their choice with the help of additional features of the experiment.[52] These additional devices changed over the years, but their main purpose – to make sure that the rational egoists choose principles that are reasonable – persisted. Thus in the 1953 Oxford notes, Rawls made the rational egoists propose principles of justice independently of each other and have these proposals be moderated by an official body.[53] The idea was that, not knowing which principles the official body will select, the rational egoists will propose principles advantageous to themselves and fair to others.[54] In the 1953-54 lectures at Cornell, this goal was achieved by the feature of the lowest representative position and the requirement that the rational egoist choose the principles of justice from that position. This feature was introduced to reflect the notion of justice, which requires "of the various institutions of society that they start from the position of assuring equal and maximum freedom to every one and depart therefrom only in such a way as to make every man better off in the long run."[55] The notions of the rational egoist and the lowest representative position were meant to serve as premises that ensure the selection of this conception of justice: "A picture of how to make a rational egoist design a just society: let him design it and give his worst enemy the option of assigning him his place in it."[56] In sum, Rawls borrowed from decision theory the simplifying devices which made his argument conclusive. The notion of reasonableness which served as the key premise of the argument did not – as we will see – come from game theory.

Rawls continued to draw links between his thought experiment and game theory in the 1950s and 1960s, acknowledging the influence but distancing his argument from the tradition's broader themes. In his 1958 article "Justice as Fairness," Rawls acknowledged the use of some of the elements of "the theory of games," but disassociated from the tradition's view of justice "as a pact between rational egoists the stability of which is dependent on a balance of power and a similarity of circumstances."[57] Unlike the theorists of games, Rawls argued, he did not commit himself to the egoist view of human motivation. From his own point of view, the resemblance of his and theory of games' approach to justice was merely superficial.[58]

---

[51] Rawls, "Justice as Fairness. Cornell Seminar," 7.
[52] Rawls, "Justice as Fairness. Cornell Seminar," "The Reasoning Game," 4.
[53] Rawls, "Oxford Notes," 49. See Rawls's description of this early version of the "original position" in his interview with *The Harvard Review of Philosophy*: Aybar, Harlan and Lee, "John Rawls: For the Record," 39-40.
[54] Rawls, "Oxford Notes," 50.
[55] Rawls, "Oxford Notes" 14.
[56] Rawls, "Justice as Fairness. Cornell Seminar," 9.
[57] Rawls, *Collected Papers*, 55-59.
[58] Rawls, *Collected Papers*, 56-57.

Only in 1965 did Rawls start attributing a greater role for the theory of games, which he then called the theory of "rational choice." In the second draft of *A Theory of Justice* [1965], where this change in self-description appears for the first time, Rawls wrote that, "if the contract theory is correct, the theory of justice belongs to the theory of rational choice. Indeed, moral philosophy is the fundamental part of this theory and of social theory generally, since ideas of right and wrong so largely define and determine human action."[59] The reasons for this change in emphasis are not clear, as Rawls's thought experiment remained the same in its essential respects. Consequently, the description of Rawls's theory as part of a rational choice should be considered as a post-hoc attempt at description – which, as Rawls himself would later accept, was a mis-description.[60]

The second version of the thought experiment was published in the 1958 article "Justice as Fairness." As its predecessor, it was part of the positivist project of showing that reasonable persons agree in their judgments. Rawls proposed to offer an "analysis of the concept of justice," or of the "principles involved in [the considered] judgments when made by competent persons upon deliberation and reflection.[61] As in 1952, he defined each feature of the experiment in order to "[bring] out a feature of the notion of justice."[62] The relevant features of justice had changed since 1952, however. In 1958 there were two of them: the constraints of having a morality and the circumstances of justice. The former implied that the principles apply to everyone equally and that no one is exempt merely because these principles are to one's disadvantage.[63] The circumstances of justice, on the other hand, depicted situations in which questions of justice typically arose. These were ones in which "conflicting claims are made upon the design of a practice and [in which] it is taken for granted that each person will insist, as far as possible, on what he considers his rights."[64]

Rawls set out to model these two features of justice in the hypothetical account and see what principles of justice they imply. The hypothetical account consisted of three key features: the persons in the original position, the problem posed to these persons, and the circumstances in which the problem has to be solved. Rawls depicted the persons in the original position as "mutually self-interested" or interested in their own objective (as opposed to relative) well-being, having "roughly similar needs and interests," and, finally, as rational.[65] Rational, in turn, implied a person who knew her interests, understood the consequences of her actions, and was capable of adhering to a plan once she decided on it.[66] Such persons had the task of choosing the principles of justice: principles to adjudicate their complaints. Rawls thought – but did not try to show – that, once circumstances of justice are imposed on this choice, the rational and mutually self-interested persons would choose the two principles of justice known as "justice as fairness."

---

[59] John Rawls, "Justice, second draft of *A Theory of Justice*. March 1965." John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 17, Folder 4, 5.
[60] Rawls, *Political Liberalism*, 53f. See also Rawls, *Justice as Fairness: A Restatement*, 82n2. Interestingly, Rawls retained the original self-description in the revised edition of *A Theory of Justice*, published in German in 1975. See John Rawls, *A Theory of Justice* (revised ed.) (Cambridge, MA: Harvard University Press, 1999), 14-15.
[61] Rawls, *Collected Papers*, 47, 71.
[62] Rawls, *Collected Papers*, 47.
[63] Rawls, *Collected Papers*, 54.
[64] Rawls, *Collected Papers*, 54.
[65] Rawls, *Collected Papers*, 52-3.
[66] Rawls, *Collected Papers*, 52.

While Rawls's thought experiment stemmed from the positivist tradition, Rawls associated it with the social contract tradition. Indeed, certain features of the positivist framework invited associations with the social contract tradition. Rawls drew the link between the positivist requirement that all reasonable persons agree in their considered judgments and the social contract requirement that the agreement to the principles of justice be unanimous. In 1958, Rawls left it ambiguous whether the social contract agreement in question took place in what he called the "general position" or among the reasonable persons in a society. Describing a fair practice, he wrote "A practice is just or fair, then, when it satisfies the principles which those who participate in it could propose to one another for mutual acceptance under the aforementioned circumstances."[67] On the one hand, this ambivalence did not matter: it was clear from the positivist requirement that the reasonable persons were expected to agree. On the other hand, however, the ambivalence was unfortunate, as it failed to specify the type of agreement – contract or the overlapping of views – which the reasonable persons in a real society were expected to reach.

This ambivalence continued in the 1962 lectures on political philosophy, in which Rawls associated the social contract tradition with both kinds of agreement. For the most part, he emphasized agreement in the general position. Explaining his own relationship to the social contract theory, Rawls wrote that he took "from the older theory (e.g. Locke) … the notion of unanimity of consent in the original position," but discarded other parts of their framework.[68] In particular, he denied that consent was historical: "That the state of nature is a historical state and one of danger etc, is left out of account in the theory."[69] Furthermore, he uprooted this consent from the background of the natural law on which the social contract theory relied; as he put it, his social contract theory was a "secularized" version of its predecessor: "The conception of justice which I shall try to work out is an elaboration of the theory of the social contract: it is a secularization of the natural rights theory [in that] … it is a natural rights theory only in the general sense, and I prefer to avoid this sense altogether."[70] Summarizing this secular social contract theory, Rawls wrote, "roughly, I want to say that an institution is just if those subject to it could have contracted into it from the original position <…>."[71] At other times in these 1962 lectures, however, Rawls emphasized agreement of reasonable persons in an actual society. "Now it follows," he wrote, "that the constitution and the basic social structure, if it is just (as defined by the two principles) can be justified to every member of the society, to every citizen."[72]

In 1964, Rawls made it reasonably clear that the social contract agreement in question was of the second kind, among reasonable persons in the actual world. As he wrote in the first draft of *A Theory of Justice*, the principles of justice "must be such that everyone could accept them and have all the knowledge of this acceptance on one another's part that they would have if they had explicitly chosen these principles."[73] In the later years he was more explicit, writing in

[67] Rawls, "Justice as Fairness," in *Collected Papers*, 59.
[68] Rawls, "Political Philosophy 171, 1962," Lecture XI, 7ii.
[69] Rawls, "Political Philosophy 171, 1962," Lecture XI, 7ii. See also Rawls, *Collected Papers*, 59.
[70] Rawls, "Political Philosophy 171, 1962," Lecture IX,1i-1ii.
[71] Rawls, "Political Philosophy 171, 1962," Lecture IX, 3i.
[72] Rawls, "Political Philosophy 171, 1962," Lecture XI, 1ii.
[73] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 69.

1970 that "Our morality is justifiable if others can be reasonably expected to accept it."[74] With this specification, the key feature of positivism – its requirement that for the objectivity of ethical judgments the agreement in the judgments of reasonable persons must be universal – became effectively re-described as a requirement of the social contract tradition. In the following sections, I will consider whether this re-description implied that Rawls abandoned his positivist framework and replaced it with the tools of the social contract tradition.

The third version of the thought experiment was introduced in the four drafts of *A Theory of Justice*, written between 1964 and 1967.[75] When these drafts were elaborated, Rawls still viewed philosophy in positivist terms. As his notes from the 1964 seminar on moral psychology show, he thought ethical theory had three tasks: those of explication, or "description by principles of the class of considered judgments," justification, or "derivation of the principles of the correct explication from philosophically defensible premises," and delineation of psychological development, or an "account of how the person comes to desire to do and to act upon what is right, to the extent that he does."[76] The thought experiment, now named the "original position," was the tool for the first two tasks of moral philosophy.

Despite the persistence of his earlier view of philosophy, Rawls started viewing his argument as Kantian from 1964 onwards. The early drafts of *A Theory of Justice* are in fact unmistakable steps in the Kantian direction; indeed, they were among the first and distinct steps in the resurgence of Kantianism in the second half of the twentieth century.[77] Nonetheless, if rational choice and social contract theory could reasonably be understood to have displaced Rawls's positivism, Kantianism clearly supplemented it by providing the key premises to Rawls's argument: an interpretation of the considered judgments of justice.

Kantianism was most evident in the conception of the person, or the features and beliefs of a person relevant to the question at hand – in this case, the question of justice. Reasonable persons will have many features that make them the particular persons they are but many of them will be features that "a rational being need not have, that is, it is not a condition of his being rational [that they have these features]."[78] In the first draft of *A Theory of Justice* [1964], Rawls

---

[74] Rawls, "Philosophy 169, 1970 Part I," Lecture 1, 10ii.

[75] Four drafts of *A Theory of Justice* that are noticeably different from the published book: one in 1964 (the very first draft), two in 1965 (although, given that there is little difference between them, they can be considered as the same draft), and one in 1967. The 1964 draft can be found in John Rawls, "Essay on Justice. First Draft of *A Theory of Justice*, 1 of 2" (1964). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 17, Folder 2. The first 1965 draft can be found in John Rawls, "Justice, second draft of *A Theory of Justice*. March 1965." John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 17, Folder 4. The second 1965 draft, distributed to students, can be found at John Rawls, "Philosophy 171. Chapters on Justice. Draft of *A Theory of Justice* reproduced to students" (1965 Fall). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 18, Folder 4. The fourth draft can be found at John Rawls, "Justice as Fairness II" (1967). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 10, Folders 1-5. There are later revisions as well, and these approach *A Theory of Justice* closer and closer. See John Rawls Faculty Papers, Harvard University Archives, HUM 48, Boxes 11-12.

[76] Rawls, "Moral Psychology, 1964-65," Seminar V, 3ii.

[77] Although Kantianism was moribund in the 1930s and 1940s, there were two significant strands of Kantianism in the 1950s: Peter F. Strawson's metaphysical Kantianism, best exemplified in his *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen, 1959), and Kantian formalism in ethics, best exemplified by Richard M. Hare's *Freedom and Reason* (Oxford: Clarendon Press, 1963) and Kurt Baier's *The Moral Point of View: A Rational Basis of Ethics* (Ithaca, NY: Cornell University Press, 1958).

[78] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 75

claimed that the principles of justice could be seen as categorical imperatives since they were expressions of "one's nature as a free and equal rational being," and since they do not "depend for [their] derivation on one's having some particular desire or end."[79] Indeed, Rawls attempted to detach the principles of justice from reasons stemming from desires or ends that do not belong to the free and equal rational being. Thus, he thought that the proper –free and equal – person justifies her political and social situation without reference to her natural and social advantages or the particulars of one's conception of a good life.

This conception of the person shaped the original position by restricting reasons used by the persons therein. Using the maxim "the knowledge [we deprive] people of is knowledge they ought not to have in choosing a conception of justice," Rawls stipulated that persons in the original position forgot the nature of their complaints against the society, their position in society (rich, poor, slave, master), their natural endowments (intelligence, gender), and the particular circumstances of their society, including its social and political institutions.[80] On the other hand, the parties knew that they had a conception of the good, and that they were subject to the circumstances of justice.[81] This was done in order to make the persons in the original position "abandon any attempt to exploit one's place in society and one's good and bad fortune in the natural lottery."[82]

Rawls's Kantianism grew and sharpened in 1965, in response to the criticism of Allan Gibbard. As Gibbard pointed out in his comments on the first draft of *A Theory of Justice*, Rawls's description of the original position did not contain explicit criteria by which the persons therein evaluate rival conceptions of justice. Without such criteria, the argument from the original position was incomplete.[83] Acknowledging the force of Gibbard's criticism, in his 1968 article "Distributive Justice: Some Addenda" Rawls introduced the notion of primary goods.[84] The primary goods were "things which rational persons may be presumed to want whatever else they want," and included goods such as liberty, opportunity, income, wealth, health, educated intelligence and self-respect.[85] Given that the primary goods were needed for any worthwhile pursuit, they were not dependent on any particular conception of a good life. This abstraction from particular contextual circumstances – even if it was restricted to the details of the original position – was one of the ways in which Kantianism seeped into Rawls's positivist theory.

Impressed by the similarities of his argument to Kant's thought, Rawls gave new descriptions for some of the old features of the original position. As a result, the origins of the original position in positivism were further removed from sight. For instance, if in the 1950s Rawls argued that mutual self-interestedness of the persons in the original position models the conditions in which questions of justice arise, by 1965 he argued that mutual self-interestedness also reflects the Kantian conception of autonomy. Mutual self-interest "means simply that persons have their own and normally conflicting systems of desires which they want to pursue,"

---

[79] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 74.
[80] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 50.
[81] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 50-53, 60.
[82] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 57.
[83] Rawls thanks Allan Gibbard for mentioning this shortcoming in the introduction to *A Theory of Justice*, and explains that the outcome of Gibbard's comments was the introduction of the notion of primary goods.
[84] Rawls, *Collected Papers*, 154-175.
[85] Rawls, *Collected Papers*, 158.

he wrote, and it is a requirement of Kantian autonomy that that one allows "perfect freedom in people's choice of their systems of desires."[86]

Similarly, Rawls now gave a Kantian interpretation to his earlier positivist conception of objectivity:

> Thus I mean by justifiable very much what Kant meant by objective. For him a principle is objective if there are reasons sufficient to determine every rational man to act on it (assuming reason to determine his will).[87]

In the 1970 lectures, Rawls similarly argued that the principles of justice apply to everyone in virtue of their being a person, regardless of their nationality or beliefs.[88] Rawls did not think that all reasonable persons would equally agree on all parts of the proposed theory. Rather, he thought, they would accept a conception of justice in different degrees:

> I disagree with Kant, I think, in that I doubt that it can be shown that there is one complete set of ethical principles in regard to which there are sufficient reasons why all rational men, given the circumstances of human life, should accept them. There may be certain objective principles in this sense, e.g., the principle not to inflict unnecessary suffering; but this principle does not make up a complete system, although it belongs to any plausible system. Indeed, I think there are degrees of justification in that certain parts of morality are more justifiable, more objective, than others.[89]

For instance, as Rawls acknowledged in 1967, he expected little agreement about the principles of distributive justice. For this reason, he allowed that the argument from the original position may not be decisive in all cases – it may exclude some feasible conceptions of justice but not others: "No doubt we shall be left with several plausible alternatives, at least on such matters as income distribution; and then we should follow the principle of tolerance: to press for the morality we favor within the limits of equal liberty." [90] Despite this qualification, Rawls's theory was expressly universal, expecting all reasonable persons to agree with the explication of their

---

[86] Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 76.

[87] Rawls, "Analytic Ethics," "1968 Ethics 169," Lecture on Justification, 1ii-2i. See also Rawls, "Philosophy 169,1970 Part II," Lecture V, 1i, where he states the problem of justification in terms of acceptability by reasonable persons: "The problem of justification: the problem of setting out reasons sufficient to convince any reasonable man that he should accept (or at least recognize that it is reasonable for us to accept) the principles that match our considered judgments in reflective equilibrium." See also Rawls, "Analytic Ethics," "1966 Philosophy 169," 5i. See similarly universalistic remarks about the concept of 'right' in Rawls, "Moral Psychology, 1964-65," Seminar VI, 3i and in Rawls, "Analytic Ethics," "1967 Ethics 169," "On justification" 1ii. For the clearest statement about universality see Rawls, "Political Philosophy 171, 1962," Lecture IX, 6i-6ii where Rawls describes the reasoning behind the formal condition of universality: "moral principles are universal: they apply to persons in virtue of their nature as human persons and not (as the law does) in virtue of their living in a certain territory or holding a certain social position."

[88] Rawls, "Philosophy 169, 1970 Part II," Lecture VII: the principles apply to everyone in virtue of being a person 6i; also Appendix to this same lecture, 2i. For a similar remark, see Rawls, "First Draft of *A Theory of Justice*, 1 of 2," 9i, where Rawls claims that the conception of justice applies to "all possible societies."

[89] Rawls, "Analytic Ethics," "1968 Ethics 169," Lecture on Justification, 1ii-2i. See also Rawls, "Philosophy 169,1970 Part I," Lecture IV, 1i-1ii.

[90] Rawls, "Analytic Ethics," "1967 Ethics 169," "On Justification," 6ii.

considered ethical judgments. Having started in positivism, this universalism now also had Kantian reasons in its support.

Despite all the new influences and new descriptions of the original position, the core of the thought experiment remained remarkably consistent from 1954 to 1967: it was meant to serve as an analysis of the considered judgments of justice. Reflections on Quine's view led Rawls to emphasize the rejection of reductionist arguments in ethics and bring to light the fact that principles of justice are justified by many kinds of considerations. The role of the original position was defined accordingly: it was to be a modeling device for this variety of considerations relevant for questions of justice. A pinnacle of the positivist argument, the original position would become the key element of *A Theory of Justice*.

7

A Theory of Justice

A Theory of Justice was an impressive combination of the topics considered over the course of twenty five years. Considerations about the subject of justice, the nature of social practices, the truths of moral psychology, the implications of having a morality, and, most of all, the nature of philosophy in ethics – all these topics were combined to carve out the scope of the argument and to support it. The conception of philosophy guiding the book was expectedly anti-foundational. As *A Theory of Justice* contained many of the same arguments made in the 1960s, I will only briefly mention them here.

Rawls rejected foundational approaches to political philosophy, claiming that "while some moral principles may seem natural and even obvious, there are great obstacles to maintaining that they are necessarily true, or even to explaining what is meant by this."[1] These considerations led Rawls to reject foundationalism entirely, as, he thought, "there is no set of conditions or first principles that can be plausibly claimed to be necessary or definitive of morality and thereby especially suited to carry the burden of justification."[2] His own approach to questions of justice was correspondingly anti-foundational, relying on the belief that conceptions of justice are supported by many kinds of considerations, that they are justified as wholes, and that they are justified not absolutely but relatively one to another.[3]

Most of all, Rawls's approach to philosophy was positivist, however much this positivism was now re-described via the social contract tropes. Rawls's argument in *A Theory of Justice* was positivist in two respects: it understood philosophy as an analysis or explication of considered judgments of justice and expected all reasonable persons to agree in a sufficient number of their judgments of justice. In other words, Rawls's anti-foundationalism was accompanied by meaning holism, but this meaning holism – like Quine's – was limited and assumed that the conceptual frameworks of all reasonable persons are sufficiently similar. This limited holism is seen in the particulars of Rawls's philosophical approach in *A Theory of Justice*. His central belief was that "justification proceeds from what all parties to the discussion hold in common."[4] As a result, his aim in the book was to gather "widely accepted but weak premises" and show that, once combined, these assumptions imply a single conception of justice or at least "impose significant bounds on acceptable [conceptions] of justice."[5] The idea was to take as premises considerations mentioned in this and earlier chapters: considered judgments most broadly and the

[1] Rawls, *A Theory of Justice*, 578.
[2] Rawls, *A Theory of Justice*, 578.
[3] Rawls, *A Theory of Justice*, 21, 579.
[4] Rawls, *A Theory of Justice*, 580.
[5] Rawls, *A Theory of Justice*, 18.

"provisional fixed points" or judgments "which we presume any conception of justice must fit" more specifically, the truths of moral psychology, and the implications of having a morality.[6]

The structure of *A Theory of Justice* reflects Rawls's beliefs about the demands of justification. The first part of the book contains the deduction of the principles of justice from considered judgments, formal constraints on the concept of justice, the implications of having a morality and a conception of the good. The second part of the book is a demonstration that the principles of justice do indeed explicate our considered judgments in particular cases and clarify the more difficult cases. Finally, the third part of the book shows that the principles of justice are consistent with the laws of moral psychology. I will concentrate exclusively on the first part of the book, since the key premise of Rawls's argument – the positivist expectation that all reasonable persons will agree in their judgments – is best exemplified therein, and since it is precisely this premise that accounts for the dilemmas of *A Theory of Justice* and therefore the subsequent development of Rawls's thought.

Rawls's goal was to use an "analytic construction," or a thought experiment to make "vivid to ourselves the restrictions that it seems reasonable to impose on arguments for principles of justice."[7] As in its previous versions, the thought experiment consisted of a chooser, the circumstances of choice, and a list of alternatives. Each particular description of the thought experiment was meant to reflect considerations relevant to questions of justice. Rawls emphasized this feature of the experiment: "Each aspect of the contractual situation can be given supporting grounds."[8] Rawls's goal was to argue that, given this defensible description of the situation of choice, two principles of justice, known as "justice as fairness," would be the unique solution to the problem of choice.[9]

The analytic construction was shaped mostly by a conception the person, which was responsible for the description of the chooser and the considerations in terms of which she chose the principles of justice. As in the mid-1960s, this conception of the person was distinctly Kantian, only this time more consciously centered on the Kantian conception of autonomy, or acting as a rational person. As Rawls wrote, an individual acts "autonomously when the principles of his action are chosen by him as the most adequate possible expression of his nature as a free and equal rational being. The principles he acts upon are not adopted because of his social position or natural endowments, or in view of the particular kind of society in which he lives or the specific things that he happens to want."[10] The conception of the rational person was unchanged: she was a "moral person," or a person with the capacities to form conceptions of the good and a sense of justice.[11] This double-edged capacity made human beings into "free and equal rational being[s]."[12] Many other features of the rational person were "the outcome of natural chance or the contingency of social circumstances" – they were not essential to being a

[6] Rawls, *A Theory of Justice*, 20. For Rawls's explicit references to such fixed points, see *A Theory of Justice*, pp. 104, 206, 311.
[7] Rawls, *A Theory of Justice*, 18.
[8] Rawls, *A Theory of Justice*, 21.
[9] Rawls, *A Theory of Justice*, 119.
[10] Rawls, *A Theory of Justice*, 252.
[11] Rawls, *A Theory of Justice*, 12.
[12] Rawls, *A Theory of Justice*, 252.

rational person.[13] These features included the individual's social position, natural endowments, the kind of society in which he lives or the "specific things that he happens to want," and other characteristics such as race and gender.[14] Justifying one's principles of justice with resort to these kinds of facts would be to lose one's autonomy and act heteronomously. Rawls's goal was to provide an argument which selects the principles of justice without endangering one's autonomy.

This Kantian conception of the person determined considerations in terms of which the persons in the original position chose principles of justice. It gave reasons – the primary goods – to evaluate alternative conceptions of justice. The content of the primary goods was determined by Rawls's understanding of a rational person, or a person capable of forming a conception of the good and developing a sense of justice: they were goods that any person needed to develop and exercise these two capacities. In short, they were goods necessary to any person as a rational person. Rawls's assumption was that rational persons in the real world prefer more primary goods rather than less; this assumption was transferred to the original position, thereby solving the dilemma of providing criteria of choice without falling into the trap of heteronomy.[15]

In the same manner, the Kantian conception of the person excluded considerations irrelevant to questions of justice. The key tool for this purpose was the "veil of ignorance," blinding persons in the original position from certain kinds of knowledge and thereby preventing them from using certain kinds of reasons in the choice of the principles of justice. Consistently with the Kantian conception of the person, the persons in the original position did not have any knowledge of the particularities of their own person, including their place in society, class position or social status, natural assets and abilities, such as intelligence and strength, or their own beliefs about the good life.[16] Nor did persons in the original position know any particular facts about their own society or the generation to which they belonged. Deliberations about justice were to be carried out without recourse to these kinds of facts.

There is no need to describe the rest of the analytic construction, or go through Rawls's argument leading from these premises to the selection of justice as fairness. But it needs to be emphasized that the original position was an integral part of Rawls's positivist approach to philosophy: a thought experiment, it was meant to gather the widely made considered judgments and, combining their force, reveal the conception of justice implicit in these considered judgments. From 1965 onward, these "widely accepted but weak premises" were becoming increasingly Kantian, and the "reasonable person" became increasingly co-extensive with the "Kantian person." These Kantian assumptions affected the content of the principles of justice, but they did not change Rawls's conception of philosophy, which remained positivist.

[13] Rawls, *A Theory of Justice*, 12.
[14] Rawls, *A Theory of Justice*, 252.
[15] Rawls, *A Theory of Justice*, 142.
[16] Rawls, *A Theory of Justice*, 137.

Rival Interpretations

I have argued that Rawls's innovation was to bring the positivist approach to bear on ethical and political questions. This interpretation is novel; in fact, Rawls is more often thought to have broken with positivism, bringing rational choice to ethics or revitalizing the social contract and Kantian traditions. These more common interpretations are supported by, and spring partly from, Rawls's own self-descriptions as a rational choice theorist, a social contract theorist, and a Kantian. This lack of correspondence between my narrative and popular narratives stems partly from the fact that positivism and the mentioned traditions are on different levels: social contract theory, rational choice theory and Kantianism do not always offer a conception of philosophy as I have understood this term in the dissertation. Thus, I believe that Rawls was both a positivist and a Kantian, and that he used the tools of rational choice theory to make his argument. In such cases, the novelty of my narrative stems from the novelty of issues I raise, not from the disagreement with other interpreters. This explanation of the divergence between my narrative and the traditional understandings is true of many cases but not all. Understood in certain ways, rational choice theory, social contract theory and Kantianism do provide conceptions of philosophy that rival that of positivism. I have assumed in the sections above that Rawls appropriated these traditions without ceasing to be a positivist in his conception of philosophy. Here I would like to make this claim good by considering arguments to the contrary and showing where they have gone wrong.

In the years immediately following the publication of *A Theory of Justice*, Rawls was often interpreted as a rational choice theorist. Robert Wolff, for instance, argued that Rawls's intention was to use the tools of rational choice to "derive substantive principles from premises that, though not purely formal, are not manifestly material either."[17] Wolff's idea captures well the essential goal of the rational choice approach to ethics and politics: to show that ethical or political principles are derivable from non-ethical, typically egoistic assumptions.[18] Its key question "Why should I do what is right?" is typically followed by an answer "Because doing what is right is in your self-interest." Rawls gave some grounds for this interpretation of *A Theory of Justice* by describing his argument as "a part, perhaps the most significant part, of the theory of rational choice."[19]

In fact, however, the relationship between rational choice theory and Rawls's argument is the inverse. Rawls acknowledged this more than twenty years later, deeming his original self-description as "simply incorrect" and stating that in fact the rational choice theory "is itself part of a political conception of justice" because "the account of the parties [in the original position], and of their reasoning, uses the theory of rational decision."[20] This later self-description is

---

[17]Wolff, *Understanding Rawls*, 20.

[18] For an excellent discussion of this approach, see Charles Larmore, *The Autonomy of Morality* (Cambridge: Cambridge University Press, 2008), 90-103. For representative works, see Richard Bevan Braithwaite, *Theory of Games as a Tool for the Moral Philosopher* (Cambridge: Cambridge University Press, 1955); Robert Axelrod, "An Evolutionary Approach to Norms," *The American Political Science Review* 80 (1986): 1095-1111; David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1987).

[19] Rawls, *A Theory of Justice*, 16.

[20] Rawls, *Political Liberalism*, 53f. See also John Rawls, *Justice as Fairness: A Restatement* (Cambridge, MA: Harvard University Press, 2001), 82n2. Interestingly, Rawls retained the original self-description in the revised edition of *A Theory of Justice*, published in German in 1975. See Rawls, *A Theory of Justice* (revised ed.), 14-15.

entirely supported by my narrative. Rawls was impressed by the deductive nature of decision theory and its consequent decisiveness. This decisiveness resulted from a clearly – and sufficiently robustly – defined chooser and the situation of choice; and Rawls intended to achieve this decisiveness by defining the situation of choice by premises sufficiently robust to yield a unique conclusion. As he wrote in *A Theory of Justice*, "The argument aims eventually to be strictly deductive. … Unhappily the reasoning I shall give will fall short of this, since it is highly intuitive throughout. Yet it is essential to have in mind the ideal one would like to achieve."[21]

The second similarity between Rawls's argument and rational choice theory is the similarity between the key definitions in the original position – rationality and mutual self-interestedness – and the key definition of rational choice theory – rationality and egoism. Rawls deliberately adopted a standard definition of rationality, also shared by rational choice theory, as "taking the most effective means to given ends" in order to "avoid introducing into it any controversial ethical elements."[22] Yet the similarity ends here. While it is true that persons in the original position are defined as mutually self-interested, Rawls's motivations for this description are different from the rational choice description of egoism. The latter definition is meant to capture the true and hidden nature of human beings.[23] Mutual self-interestedness, on the contrary, is meant to reflect the implications of the concept of morality, or what Rawls called the "circumstances of justice."[24] When questions of justice arise, Rawls thought, they arise because persons advance conflicting claims to social goods and are unwilling to concede their position on reasons other than those relevant to justice.[25] Thus sympathy, pity and other irrelevant reasons do not make the claimants cede their claims. The ground for this description is that, "as a matter of realism, this is how things are," although it would be more appropriate to say that questions of justice should be decided only by reasons relevant to such questions.[26] The condition of mutual disinterestedness of the parties in the original position was meant to reflect this feature of the circumstances of justice and ensure that the choice of principles does not depend on sentiment and affection. As Rawls summarized this condition in his 1962 lectures on political philosophy, "the point of the 'mutually' [in 'mutually self-interested'] is only to indicate that the parties are not self-interested simpliciter (they are not rational egoists), but they regard themselves as having legitimate interests which they are prepared to press on one another…."[27] In sum, then, Rawls used rational choice theory as a useful guide for the analysis of our conception of justice, but never saw it as a correct conception of philosophy. Rational choice theory had very different implications in ethics; Rawls admired and adopted some aspects of this approach, but rejected its broader aims.

Rawls has often rightly been understood as a Kantian. Rawls himself called justice as fairness "highly Kantian in nature," and stated that "there is a Kantian interpretation of the

---

[21] Rawls, *A Theory of Justice*, 121. See also his remarks in "Justice as Reciprocity," an article published in the same year at *A Theory of Justice*. John Rawls, *Collected Papers*, 203-4.
[22] Rawls, *A Theory of Justice*, 14.
[23] See, for example, Edgar Kiser and Michael Hechter, "The Debate on Historical Sociology: Rational Choice Theory and Its Critics," *American Journal of Sociology* 104 (1998): 785-816, esp. p.802.
[24] Rawls, *A Theory of Justice*, 126-130.
[25] Rawls, *A Theory of Justice*, 129-130.
[26] Rawls, "First draft of *A Theory of Justice*," 75.
[27] Rawls, "Political Philosophy 171, 1962," Lecture X, 1ii-2i.

conception of justice from which the principles derive."[28] Rawls's self-descriptions in this case are entirely accurate: as we have seen, the conception of the person on which the argument in the original position relies is highly Kantian. Yet, despite the Kantian aspects of his principles of justice, Rawls was not a Kantian in his conception of philosophy: he did not justify his principles of justice in a Kantian way, however broadly conceived. This can be best seen in the failures of the contrary argument, recently made by Robert S. Taylor. According to Taylor, Rawls's Kantian conception of the person is a "necessary presupposition or postulate of practical reason."[29] In Kant's theory, as Taylor explained, this conception of the person is established either directly, by showing that it is "something we must presuppose if we are to conceive of ourselves as agents, which is unavoidable," or indirectly, by showing that it is presupposed in the "fact of reason" that makes us conscious of our freedom.[30] In Kant's theory, this conception of the person is therefore a "self-evident first principle," which implies, as Taylor goes on to show, that it cannot be rejected if it contradicts some of our considered judgments.[31] Thus Kantianism as a conception of philosophy is characterized by two key features. First, it is an attempt to derive ethical conclusions from considerations about what it is to make an ethical judgment, or, more broadly, what it is to take the standpoint of practical reason. Second, these conclusions are viewed as necessary given that taking the practical standpoint is unavoidable.

However illuminating this conception of Kantianism is of 20th century ethical thought, it is not a fitting description of Rawls's approach to philosophy.[32] Two reasons stand out. First, Rawls has always understood moral philosophy as analysis of considered judgments, and, in the anti-foundational way which he made very explicit in the 1960s, he allowed that, in principle, any considered judgment – as any part of a moral theory – can be rejected as misguided. He reaffirmed this anti-foundationalism in *A Theory of Justice*, emphasizing that "even the judgments we take provisionally as fixed points are liable to revision."[33] As part of these considered judgments, the Kantian conception of the person is also in principle liable to revision, even if in practice Rawls was confident that it describes the considered judgments correctly. This way of arriving at the conception of the person is clearly incompatible with Kant's: Rawls did not claim that this conception of the person was in any way self-evident or necessary. In Rawls's own terms this would have been a Cartesian move. Instead, he sidelined the concept of "necessity" altogether, claiming that without a broader background in which "necessity" acquires philosophical significance, this concept has no use.

Neither is Rawls's conception of philosophy Kantian in the second respect: principles of justice are not defended as implications – whether these implications are necessary or not – of practical reason. Rawls disowned this interpretation of his later arguments in *The Law of Peoples*, stating explicitly that "at no point are we deducing the principles of right and justice … from a conception of practical reason in the background."[34] While no such explicit statement can be found in *A Theory of Justice*, Rawls's 1999 disassociation from the Kantian argument applies

[28] Rawls, *A Theory of Justice*, viii and 251, respectively. See especially Ibid., 251-257.
[29] Taylor, *Reconstructing Rawls*, 234.
[30] Taylor, *Reconstructing Rawls*, 234.
[31] Taylor, *Reconstructing Rawls*, 234-5.
[32] This conception of Kantianism describes well the main goals of Korsgaard, Herman, and Onora O'Neill. See Korsgaard, *Sources of Normativity*; Herman, *Moral Judgment*; O'Neill, *Justice and Virtue*.
[33] Rawls, *A Theory of Justice*, 20.
[34] Rawls, *Law of Peoples*, 86.

fully to this argument in 1971. Admittedly, Rawls did draw implications of making an ethical judgment in *A Theory of Justice*: these were the constraints on the concept of right, including universality and finality. If one makes an ethical judgment, Rawls assumed, it applies to all persons in similar conditions and cannot be changed if it goes against one's interest. These constraints on the concept of right were incorporated into the argument from the original position, but only as part of the many considerations required to deduce principles of justice. Throughout the 1960s and in *A Theory of Justice*, Rawls maintained that the principles of justice cannot be derived from any one kind of consideration. In particular, he thought, principles of justice could not be derived from formal conditions on the concept of right. As we have seen in Chapter Five, Rawls criticized Kurt Baier and Richard Hare precisely for trying to derive principles of morality from the conditions of prescribability and universalizability.[35] Rawls argued then that "we cannot … derive the content from the formal conditions alone": "this is too slender a basis."[36] Thus, even understood in this Rawlsian way, the practical standpoint played only a partial role in Rawls's argument for principles of justice. In sum, then, *A Theory of Justice* was clearly Kantian in its conception of the person and the content of the principles of justice. But the conception of philosophy driving the book was not Kantian: it did not treat ethical principles as implications of making an ethical judgment, much less as necessary implications.

Rawls has also been interpreted as a social contract theorist, and is still often seen as such today. This interpretation persists in virtue of Rawls's requirement that a conception of justice be universally acceptable. As Rawls required both that all persons in the original position agree on the principles of justice and that all reasonable persons in the real society do so, the requirement of universal acceptance has spawned two main interpretations of Rawls as a social contract theorist. The first emphasized agreement in the original position while the second stressed agreement in the actual society. Yet neither of these interpretations account for Rawls's main goals in *A Theory of Justice*. The first mistakenly places the emphasis on the agreement in the original position. The second has many virtues: it rightly concentrates on the agreement among reasonable persons in the real world and insightfully points to the presence of the social contract elements in Rawls's argument. Nonetheless, it too fails to note that the contractualist requirement does not introduce anything new to Rawls's conception of philosophy; indeed, it imposes itself upon the already existing the positivist expectation that all reasonable persons would agree in their considered judgments.

Rawls gave grounds for the social contract interpretation of his work by describing *A Theory of Justice* as the "traditional theory of the social contract" but "generaliz[ed] and carri[ed] to a higher order of abstraction."[37] A year later, in a 1972 discussion with Stanley Moore, he further described his argument as merely an extension of Jean-Jacques Rousseau's *Social Contract*'s core insights. Key among these was Rousseau's claim that the general will is universal in its source – it is shared by all: "the general will, to be truly what it is, must be

[35] For Rawls's description of Hare as a formalist, see Rawls, "Philosophy 169, 1970 Part II," Lecture IX, 4i. Rawls also includes Jonathan Harrison and Henry D.Aiken among formalists; see Ibid, 5ii. For Rawls's classification of Baier as a formalist, see more broadly Ibid., Lecture VII.
[36] Rawls, "Philosophy 169, 1970 Part II," Lecture VIII, 3b. See also Rawls, *A Theory of Justice*, 251, 51.
[37] Rawls, *A Theory of Justice*, viii.

general in its purpose as well as in its nature; … it should spring from all and apply to all."[38] As Rawls wrote, this passage fully describes the essence of his arguments in *A Theory of Justice*:

> that passage [of Rousseau] had a great effect on me. I can first recall reading it (at least with understanding) around 1958. By that time the fundamental intuitive idea of A Theory of Justice had long since occurred to me (1950-51), and I had already thought about many of the problems in trying to work it out. With this conception in mind, I was ready to grasp the significance of what Rousseau was saying. The discovery of Rousseau finally dispelled any pretense of originality for the idea I had been thinking about; and led me to recognize that the essential thing was to develop the contract doctrine into a reasonably clear moral theory.[39]

Commentators have agreed with Rawls's self-description, but they differed in interpreting the way in which the principles of justice came from all and applied to all. Some, notably Ronald Dworkin, claimed that the principles of justice derive their legitimacy both from the fact of their universal acceptance in the real world and from the fact of universal agreement in the original position.[40] The text of *A Theory of Justice* provides support for this interpretation, as Rawls linked the social contract theory to the requirement that the principles of justice be "the object of the original agreement," when this agreement takes place "in an initial position of equality as defining the fundamental terms of their association."[41] The initial position of equality in this context referred to the original position. Dworkin then went on to criticize this argument, claiming that, since the agreement in the original position is hypothetical, it cannot justify the principles of justice.

Rawls disowned this social contract interpretation of *A Theory of Justice*, arguing that the original position, and therefore the contract that takes place in it, is a thought experiment meant to model the considered judgments of reasonable persons in the real world.[42] As he wrote already in 1964, explaining the grounds for political obligation, the argument in the thought experiment drew its force not from the feature of contract, but from the considerations relevant to the question at hand:

> But now in all this [the argument for the principles of justice] no reference is made to an actual agreement. It is not because an actual agreement has been made that any one is bound; but because certain acts etc have the requisite property of being in accordance with such and such principles. The fact that in formulating this property an analytic hypothetical contractual model is introduced is irrelevant. After all, it is logically possible in advance that one would agree to have no promises; that there is a contractual model doesn't in itself account for

[38] Jean-Jacques Rousseau, *The Social Contract*, trans. Maurice Cranston (London: Penguin, 1979), 75.
[39] John Rawls, "Reply to Stanley Moore" [1972]. John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 19, Folder 4, 1i.
[40] Ronald Dworkin, "The Original Position," *The University of Chicago Law Review* 40 (1973): 500-533.
[41] Rawls, *A Theory of Justice*, 11.
[42] John Rawls, "Justice as Fairness: Political not Metaphysical" in Rawls, *Collected Papers*, 400-401.

> the obligation of actual promises. It is the whole construction and what one can say about it [that accounts for this obligation].[43]

The goal of the original position – and therefore of the contract that takes place within it – was to reveal something about the considered judgments of reasonable persons in the actual world. The agreement among the persons in the original position was therefore also meant to reveal something about our considered judgments. If the contract in the original position further justifies the principles of justice, it is because it reveals that the premises, when combined together, are indeed strong enough to select one conception of justice and to show thereby that all reasonable persons share a sufficient number of considered judgments.

Once the agreement is said to take place among reasonable persons in the real world, not in the original position, the interpretation of *A Theory of Justice* as a work in the social contract tradition becomes far more defensible. Samuel Freeman defended this "contractarian" interpretation, claiming that "[it] is this general agreement among the members of a well-ordered society that mainly drives the contractarian element in Rawls's view…"[44] This interpretation is not only plausible but also incompatible with the one I have provided, since contractarianism understood in this way would offer a rival conception of philosophy to that of positivism. The contractarian conception of philosophy revolves around the idea of acceptability without agreement: it attempts to forge a conception of justice that is acceptable to all although not held by all. On this view, political philosophy is a practical enterprise, aimed at securing feasible agreement. Thomas M. Scanlon's distinction between a position that is one's own and a position that cannot be reasonably rejected, is a good example of contractarianism's main idea.[45] This conception of philosophy is clearly incompatible with that of positivism, which sees its task as an empirical, not practical enterprise, and expects to find agreement, not forge it.

Although there are certainly passages in *A Theory of Justice* which support the contractarian interpretation of Rawls, the core contractarian expectation that political philosophy is a practical activity involving compromises plays no role in Rawls's argument. Indeed, it plays no role precisely because of the positivist assumption that all reasonable persons already agree in their judgments and that all that is needed is to collect these judgments and analyze their implications. There is no need for a compromise along the lines described above. This can be seen from a contrast between *A Theory of Justice* and Rawls's later argument in *The Law of Peoples*. In this later book, in which Rawls ceded the key positivist assumption that all reasonable persons would agree in their judgments of justice, he argued that different reasonable people would justify the law of peoples from their own points of view, using different kinds of arguments.[46] The argument from the original position was treated as only of the possible argument for the laws of peoples. This first step toward the contractarian position – acknowledgment of the possibility of deep disagreement – is missing in *A Theory of Justice*. Every reasonable person is expected to go through the same kind of reasoning.

In sum, the history of the original position reveals Rawls's changing commitments. First developed in the mid-1950s as part of the positivist tradition, it was meant to test Rawls's main

---

[43] John Rawls, "Moral Psychology, 1964-65," Seminar VI, 7ii.
[44] Freeman, *Justice and the Social Contract*, 4.
[45] Scanlon, *What We Owe to Each Other*, 189-247.
[46] Rawls, *Law of Peoples*, 11-12, 58, 68.

assumption that all reasonable persons agree in their judgments, and that judgments of justice could thereby be declared objective. Yet as Rawls was influenced by the rational choice theory in the early 1950s, then by the social contract tradition in the late 1950s, and, finally, by Kantianism in the mid-1960s, he began to re-describe the purposes of the original position and its original connection with positivism was lost from sight. Yet despite this re-purposing of the original position, it remained as a feature of Rawls's broader positivist framework: it was meant to serve as the analysis of the considered judgments of reasonable persons. Some of the traditions that influenced Rawls were complementary to positivism: Kantianism provided premises from which the two principles of justice were derived, while rational choice theory filled in some of the reasoning by which this derivation took place. The social contract tradition, which in principle could have rivaled Rawls's positivist conception of philosophy, ended up being a vacuous re-description of the positivist requirement that all reasonable persons agree in their judgments. Rawls continued to see philosophy as an empirical inquiry, analyzing the judgments of reasonable persons and hoping to reveal their underlying unity.

Positivism's Latent Dilemmas

A Theory of Justice, a summary of Rawls's twenty-five year effort, was a work grand in its aspirations and achievements. The positivism that inspired the book made for an imposing theory of justice, in two respects: its universal scope and its simple, well-ordered, and mechanical account of ethical judgment. Yet the magnitude of this edifice hid an underlying weakness from sight: the possibility that all reasonable persons may not agree in their judgments of justice. Prior to 1971, Rawls dealt with this dilemma by claiming that, if all reasonable persons do not agree entirely, they do share a family of overlapping frameworks of thought. That was a retreat from the initial positivist picture of philosophy; in *A Theory of Justice* this retreat showed in a modified theory of judgment: the principles of justice guided reasonable persons only in a general direction of political action. The positivist ideal of mechanical judgment independent of intuition was forgone. Thus, to a person familiar with the history of Rawls's intellectual development, his argument revealed a fundamental weakness. The positivism that made for the grandeur of *A Theory of Justice* would also be the cause of its downfall.

A Theory of Justice was a grand work primarily because of its universalism: its conclusions were meant to apply to all persons insofar as they were reasonable persons, and, as "reasonable" did not carry stringent requirements, the principles in fact applied to all persons. As Rawls wrote in his 1970 lectures on ethics, the principles apply to "all moral agents (persons): those who can understand and act upon prima facie principles."[47] In 1946, when Rawls started his project, this universalism was justified as a requirement of the positivist theory: if ethical judgments were objective, they had to exhibit regularity among reasonable persons. By 1971, Rawls had given two further independent reasons for this positivist justification. First, he defended this universalism as the Kantian requirement for autonomy: the principles of justice were justified to persons insofar as they were rational persons, regardless of their social position, nationality, beliefs about the good life, or other non-essential traits. Second, he saw universalism

[47] Rawls, "Philosophy 169, 1970. Part II," Lecture VII, 2i.

as a demand of the social contract theory, since, for the principles of justice to be justified, they had to be acceptable by all reasonable persons.

Positivism made *A Theory of Justice* a grand work in another respect as well: it made it into a *theory* of justice, one that identifies reasons relevant to questions of the justice of social institutions, orders these reasons in terms of importance, and assumes that these ordered reasons should guide our judgment in *all* questions about social justice. This view was ambitious in two respects: it offered a tidy view of questions about social institutions, assuming that all such questions are decided in terms of the same reasons, and it sought to avoid reliance on intuition as far as possible by providing a "replacement schema" for our unanalyzed ordinary concepts like "justice" and "rightness." Like positivism in other areas of life, *A Theory of Justice* aimed to make the reasons for our judgments ordered and austere.[48]

But the positivism of *A Theory of Justice* was modified after the twenty five years of responding to objections and solving dilemmas. As a result, the actual achievement of the book – however grand – did not live up to the grand challenge posed to ethical theory by positivism. These modifications were responses to the problems with the key positivist assumption that all reasonable persons share a conceptual background. Throughout the years Rawls modified this assumption, and these modifications reverberated into other parts of his argument.

Rawls's attempts to deal with the weaknesses in the main positivist assumption are most seen in *A Theory of Justice*'s noticeably limited account of judgment. Already in the late 1950s, Rawls argued that all reasonable persons do not agree on every issue but rather share a family of views: disagreeing on some questions of justice, they nonetheless share large parts of the conceptual schemes and as a result make many of the same judgments. This expectation to find a shared overlap of views but not strict agreement continued into *A Theory of Justice*, where Rawls wrote:

> I have assumed that in a nearly just society there is a public acceptance of the same principles of justice. Fortunately this assumption is stronger than necessary. There can, in fact, be considerable differences in citizens' conceptions of justice provided that these conceptions lead to similar political judgments. And this is possible, since different premises can yield the same conclusion. In this case there exists what we may refer to as overlapping rather than strict consensus.[49]

Limiting the aim from showing strict agreement to showing overlapping agreement also jeopardized the goal to free the theory of justice from its dependence on intuitive – not mechanically guided – judgment. Rawls was forced to admit that a theory of justice would not direct the judgment of all reasonable persons to the same conclusion for the following reason. The principles of justice were chosen on the basis of their capacity to explicate the fixed points in our judgments of justice and some, perhaps most, considered judgments – more of such judgments, in any case, than rival theories could explicate. But the best principles did not have to explicate all considered judgments. It was not therefore possible that these principles, after having been selected, would then imply agreement on all cases of justice. In short, wanting to

[48] See also the *Aufbau* movement in architecture. Galison, "Aufbau/Bauhaus" and Galison, "Constructing Modernism."
[49] Rawls, *A Theory of Justice*, 387-8.

maintain the universalism of his theory by allowing that the principles permit a degree of disagreement, Rawls modified the grand picture of the theory and especially its unwillingness to rely on intuitive judgment. Justice as fairness was much more limited compared to the initial positivist picture: it was a "guiding framework" that directed the judgments of reasonable persons in a general direction. Rawls's theory of justice was universalistic, but it relied on a limited extent of agreement.

A Theory of Justice was a guiding framework in two ways, and it is important to distinguish between these. In the first way, it often did not provide sufficient reasons for a political decision. From the early 1960s onward Rawls presented justice as a single virtue on par analytically with other political virtues such as efficiency and liberality. Justice was not, he emphasized, a "social ideal," or a full-fledged vision of an ideal society.[50] It was, on this interpretation, "but a part, although perhaps the most important part, of such a conception."[51] Justice was a necessary condition for an acceptable social ideal: "Justice is the first virtue of social institutions …. laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust."[52] But justice was not a sufficient condition for an acceptable social ideal. And, as political decisions regarded the entire social ideal, considerations of justice did not always lead to a political decision by themselves. To decide these questions, additional considerations were required. In that regard, *A Theory of Justice* was a limited theory: it was not always decisive about all particular political questions.

Rawls noted three important cases, all of them concerned with questions of economics, in which considerations of justice were not sufficient to decide. Most famously, he acknowledged that justice as fairness did not provide sufficient considerations to decide between the socialist system of ownership and property-owning society, since the decision depends "in large part upon the traditions, institutions, and social forces of each country, and its particular historical circumstances" – matters which a theory of justice did not discuss.[53] Similarly, Rawls did not think that a theory of justice helps determine the extent of legitimate economic power or the rate of savings between different generations.[54]

Rawls did not see these limitations as a failure on his part. Rather, he thought they showed the proper limits of philosophy. As he told his students in the 1966 lectures on ethics, "moral philosophy must stop, qua moral philosophy, at the general framework. For only a careful study of the facts etc can determine what to do in particular situations."[55] Once the limits of moral philosophy are reached, the range of cases approved by the considerations of justice is to be considered equally just: "justice is to that extent likewise indeterminate. Institutions within the permitted range are equally just, meaning that they could be chosen; they are compatible with all the constraints of the theory."[56] This first way in which *A Theory of Justice* was a limited theory – and particularly its deliberately narrow understanding of justice – is interesting, but it is not

[50] Rawls, *A Theory of Justice*, 9.
[51] Rawls, *A Theory of Justice*, 9.
[52] Rawls, *A Theory of Justice*, 3.
[53] Rawls, *A Theory of Justice*, 274.
[54] Rawls, *A Theory of Justice*, 278, 287. See *A Theory of Justice*, 362 for Rawls's explanation why one cannot require precision in the case of the rate of savings.
[55] Rawls, "Analytic Ethics," "1966 Philosophy 169," "Analytic Ethics: Metaethics," 7i.
[56] Rawls, *A Theory of Justice*, 201. See also *A Theory of Justice*, 199.

telling of Rawls's positivist approach: it is possible that all reasonable persons agree in all matters of justice but nonetheless these considerations only carry so far. For this reason I will leave this first limitation aside.

The second, more relevant, way in which *A Theory of Justice* is limited is telling of the limitations of the positivist approach to philosophy. Rawls understood justice as fairness as a guiding framework in the sense that it did not always guide our intuitions in a particular direction. Rawls explained the guiding framework view in contrast to the more ambitious – and more properly positivist – view of moral theory as a "deductive schema."[57] On this latter view, moral principles serve as "the major premises of our moral judgments," whereas "facts of the case" serve as the minor premises. Together, these premises "generate our considered judgments in full reflective equilibrium."[58] A distinctive feature of such a "deductive schema" is that it does not require intuitive judgment because all reasons relevant to the question at hand weigh univocally in one direction: "no principle applies and points in another direction (supports some other alternative).[59] In such a case, "no moral decision (deliberation) is necessary. The answer is obvious."[60] Such situations, according to Rawls, are rare and select, and therefore the deductive schema view of ethical theory is fitting "only in special cases."[61]

Rawls's more "realistic and accurate" – but also more limited – conception of ethical theory was the "guiding framework" conception.[62] This conception acknowledged the possibility of conflicting reasons and admitted that the priority rules for ordering these reasons would guide our judgment in some but not all cases. The principles of justice could not solve all cases of conflict. Therefore, even the best of the feasible theories of justice "identifies the relevant considerations and helps us to assign them their correct weights" but does so only in the more important cases.[63] Consequently, the task of such a feasible theory "when addressing the priority problem … is that of reducing and not eliminating entirely the reliance on intuitive judgments."[64] Along the same lines, Rawls also compared ethical theory to economic theory "which largely tells us what to look for."[65] In short, key to the "guiding framework" conception was the acknowledgment that, despite the priority rules, in some cases even the best conception of justice requires "the exercise of some judgment [unguided by the priority rules]." [66]

Rawls did not explain why ethical theory cannot provide guidance to judgment in all cases, but his earlier engagement with Wittgenstein brings these reasons to light. Ethical theory could not set a proper ordering for all cases and for all reasonable persons because reasonable persons disagreed when placing weights on different values. This can be best seen in Rawls's 1958 Wittgensteinian investigations. Describing the limits of morality, Rawls wrote that "all

[57] Rawls, "Philosophy 169, 1970. Part II," Lecture VI, 1i.
[58] Rawls, "Philosophy 169, 1970. Part II," Lecture VI, 1i.
[59] Rawls, "Philosophy 169, 1970. Part II," Lecture VI, 6i.
[60] Rawls, "Philosophy 169, 1970. Part II," Lecture VI, 6i.
[61] Rawls, "Philosophy 169, 1970. Part II," Lecture VI, 6i-ii.
[62] Rawls, "Philosophy 169, 1970. Part II," Lecture VI, 2i.
[63] Rawls, *A Theory of Justice*, 44-45, 364. See also *A Theory of Justice*, 41.
[64] Rawls, *A Theory of Justice*, 44-45, 364. See also *A Theory of Justice*, 41.
[65] Rawls, "Philosophy 171, 1966," Lecture I, 2ii.
[66] Rawls, "Philosophy 171, 1966," Lecture I, 2ii.

moralities resemble one another in their principles; they have this sort of family likeness."[67] Resembling one another in some of their principles, these moralities differ "by varying the emphasis and so favoring one principle over another."[68] As a result, while in most cases all moralities guide their practitioners in the same direction, in some particular instances they may disagree. In such cases, however, justice is not indifferent because different moralities would consider different courses of action as just. As different reasonable persons disagree in courses of action, a theory of justice is conflicted in its recommendations. This conclusion was the price Rawls had to pay in order to maintain the universalism of his theory: willing to admit that some disagreement in judgments is admissible for principles that are universal, he was thereby also forced to admit that these universal principles will not guide all reasonable persons to the same conclusion. In short, maintaining his core assumption that all reasonable persons share a conceptual framework – now said to be a family of such frameworks – Rawls was forced to admit that a theory of justice served only as a "guiding framework" in practical political affairs.

In sum, when *A Theory of Justice* came to print in 1971, it appeared grandiose, claiming universality for its conclusions and aiming to replace the reliance on intuition by ethical principles and priority rules. Yet the argument fell short of its aims: Rawls admitted that the principles of justice could not be expected to guide the judgments of reasonable persons to the same conclusion. In practical politics, the principles played a role of a guiding framework. This limited achievement – limited in light of its aim – was not without reason. It indicated problems with the key positivist assumption that all reasonable persons agree in their conceptual framework. Rawls had scaled down this assumption to the claim that all reasonable persons shared a family of conceptual frameworks. This claim was to be tested by the readers of *A Theory of Justice*. As a result of criticisms, Rawls's argument would have to be reworked without its central positivist pillar.

[67] Rawls, "Essay V," 3: 1.
[68] Rawls, "Essay V," 3: 1.

8

Epilogue: Positivist Dilemmas, Positivist Developments

Over the years, *A Theory of Justice* has become one of the most cited works of 20th century political philosophy. Translated into over twenty languages, it has originated or reinvigorated political traditions, yet it has also received a number of important criticisms. In this Epilogue, I want to give a brief overview of the evolution of Rawls's project, focusing on the early criticisms it received. These criticisms brought the dilemmas implicit in his positivist approach philosophy to Rawls's attention and prompted important changes in his argument.

Resting on the key positivist assumption that all reasonable persons share a conceptual framework, the argument of *A Theory of Justice* consisted in gathering "widely accepted but weak premises" and showing that, taken together, these premises imply justice as fairness.[69] An argument that rests on shared assumptions is most vulnerable to claims that these assumptions are not actually shared, and, indeed, such were the most common objections to Rawls's argument. As the allegedly shared premises were modeled in the original position, this device took the brunt of such objections.[70]

The most potent of such criticisms was the claim that Rawls's argument relied on a conception of the person that was not, as intended, widely accepted or weak.[71] Others have similarly questioned Rawls's reliance on Jean Piaget and Lawrence Kohlberg's laws of moral psychology, as these thinkers were self-avowedly Kantian and started their arguments from premises which utilitarians did not accept. Similarly, critics contested Rawls's reliance on the allegedly widely shared scientific knowledge which was made available to the persons in the original position.[72] This knowledge, they claimed, is contested and cannot be incorporated into the "widely shared" premises.[73]

In essence, critics of *A Theory of Justice* brought to light how particular – and sometimes odd – Rawls's premises were. In doing so, they challenged his central assumption that reasonable persons share a conceptual framework – or, at least, a family of overlapping conceptual frameworks. Instead, the critics pointed out, the premises which characterized the original

---

[69] Rawls, *A Theory of Justice*, 18.

[70] Of course, Rawls's argument contained the caveat that the premises have to be shared or become shared "after reflection," which includes going through the process of reflective equilibrium. As such, the argument is not damaged by the mere fact of the contestation of premises; some process of reflection and arguing back and forth must be allowed to take place. Yet the caveat "after reflection" cannot work as an indefinite escape, holding the hope of eventual agreement of all reasonable persons far in the horizon.

[71] Michael J. Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982).

[72] See Leon H. Craig, "Contra Contract: A Brief against John Rawls' "Theory of Justice"," *Canadian Journal of Political Science* 8 (1975): 68-71.

[73] For a well-known criticism of Kohlberg's laws of moral development, see Carol Gilligan, *In A Different Voice* (Cambridge, MA: Harvard University Press, 1982).

position are contested and, moreover, interconnected. Increasingly, the critics were portraying *A Theory of Justice* as a work that is consistent, impressive in its scope but nonetheless particular – Kantian. They stressed that other frameworks of thought would disagree about the description of the original position because they relied on a different set of premises. As a result – these critics implied – a further defense of Rawls's Kantian judgments and the three laws of moral psychology was required.

Rawls's initial response to such criticisms was to defend the key positivist assumption by denying the extent of meaning holism. He started claiming that the comparison of theories of justice can take place regardless of other areas of inquiry. Thus, in the 1975 article "The Independence of Moral Theory," Rawls argued that the study of considered judgments, although not isolated, is very much an inquiry independent of other types of inquiries: "much of moral theory is independent from the other parts of philosophy. The theory of meaning and epistemology, metaphysics and the philosophy of mind, can often contribute very little [to questions raised by moral theory]."[74] In particular, Rawls denied the links between moral theory and views about the conception of the person: "the conclusions of the philosophy of mind regarding the question of personal identity do not provide grounds for accepting one of the leading moral conceptions rather than another."[75] In short, as the critics argued that the Kantian conception of the person depended on commitments in other areas of inquiry, including his reliance on the Kantian laws of moral psychology, Rawls denied such connections. Initially, then, he believed that the truth of each aspect of his argument could be established within the confines of its own domain, be it in moral theory or in philosophy of mind. He reaffirmed the 1946 positivist stance, claiming that moral theory has its own subject matter, although by 1975 this subject matter expanded to include more than considered judgments of reasonable persons: "the study of substantive moral conceptions and their relation to our moral sensibility has its own distinctive problems and subject matter that requires to be investigated for its own sake."[76]

Within five years of publishing "The Independence of Moral Theory," however, Rawls had given up the defense of the positivist severance of the description of the original position from the deeper commitments on which this description relied. On Rawls's own account, Samuel Scheffler's "Moral Independence and the Original Position," published in 1979 but sent to Rawls in 1977, played a crucial role: it is then that he realized the need to significantly revise the argument of *A Theory of Justice*.[77] Scheffler's article pointed out an inconsistency between the argument in the original position and Rawls's claim that moral theory is independent of other areas of inquiry. In its broadest claim, however, Scheffler's argument was illustrative of the kinds of criticism Rawls's theory had already received: it pointed out that Rawls's argument relied on deeper commitments which rival theories of justice did not share. Yet it must have been more influential than other arguments because it showed precisely how Rawls's argument against rival conceptions of justice in the original position depended on the deeper commitments.

---

[74] Rawls, *Collected Papers*, 286.
[75] Rawls, *Collected Papers*, 296.
[76] Rawls, *Collected Papers*, 287.
[77] Samuel Scheffler, "Moral Independence and the Original Position," *Philosophical Studies* 35 (1979): 397-403. For Rawls's remarks about the paper, see Rawls, *Political Liberalism*, xxxii-xxxiii.

Scheffler argued that Rawls's argument against utilitarianism depended on the Kantian conception of the person – and so was not independent of other fields of inquiry, as Rawls claimed in the 1975 article. Conceiving herself in terms of long-term life plans and interests, Rawls's Kantian person rejected the utilitarianism which endangered these long-term interests by permitting inadmissible sacrifices of individual liberties for the common good. But, Scheffler insisted, utilitarianism did not accept the Kantian conception of the person, allowing that a person may in fact be a bundle of immediate desires without the long-term plans and interests. Thus, Rawls's argument against utilitarianism, insofar as it was successful, drew its force from the fact that its implications on the questions of personal identity were more defensible than those of utilitarianism.

By the late 1970s, Rawls increasingly realized the depth and extent of potential disagreement among reasonable persons and acknowledged that disagreement about one subject often leads to disagreement about ethical theories. As a result of this realization, Rawls concluded that rival ethical theories may not have a sufficient number of weak and widely accepted premises to deduce a conception of justice. This created a dilemma for Rawls: his central positivist assumption that all reasonable persons share a conceptual framework – or a family of overlapping frameworks – seemed flawed.

Rawls's subsequent intellectual development can be best understood as a response to this dilemma. Acknowledging that his central assumption was flawed, he started reorganizing his theory without this assumption. His new defense of justice as fairness drew on the positivist themes: he argued that, despite broader disagreements, reasonable persons shared a political culture. His argument consequently became more contextual and historical, but the contextualism and historicism were limited by the key positivist themes.

Rawls's solution to the dilemma in his positivist framework was twofold. First, he sharply distinguished between different parts of our web of beliefs: the comprehensive and political doctrines. This distinction consisted in demarcating our broader beliefs about the world, intellectual inquiry, and the nature of morality from beliefs about the political sphere alone.[78] The second step was to acknowledge the fact of reasonable disagreement in the comprehensive sphere, but posit the fact of agreement in the public sphere – the existence of a "public political culture" – and then attempt to formulate a conception of justice from these shared public beliefs. As he wrote in 1985, "We look, then, to our public political culture itself, including its main institutions and the historical traditions of their interpretation, as the shared fund of implicitly recognized basic ideas and principles. The hope is that these ideas and principles can be formulated clearly enough to be combined into a conception of political justice congenial to our most firmly held convictions."[79]

The key novelty in this argument was the requirement that each different comprehensive doctrine justify the political conception of justice on its own grounds, from within its own broader conceptual framework. Rawls's argument thereby became more contextual: it now acknowledged the different and incompatible starting points of the rival comprehensive

[78] See, for example, John Rawls, "Justice as Fairness: Political Not Metaphysical," (1985) in Rawls, *Collected Papers*, 388-414 and Rawls, *Political Liberalism*, 1-15.
[79] Rawls, *Collected Papers*, 393.

frameworks. The argument from the original position consequently became treated as an argument only from a comprehensive liberal point of view, not relevant to those rejecting the liberal comprehensive premises.

Despite the new contextualism, Rawls's new argument remained remarkably positivist. Although Rawls no longer believed that all reasonable persons shared a comprehensive framework, he continued to hold that their beliefs about the political sphere would overlap sufficiently to inform a political conception of justice. In essence, then, Rawls retained his earlier positivist belief that the extent of meaning holism does not threaten the agreement among reasonable persons. He did not think that disagreement about comprehensive frameworks would extend to political doctrines. In that key regard, Rawls's later argument remained indebted to positivism.

Rawls's later work, thus, is a development of the positivist framework expounded in *A Theory of Justice*. This later work is therefore a further test of how fruitful a modified positivism in ethics can be. The new positivism brings the sensitivity to disagreement further, but it stops short of abandoning the hope that all reasonable persons agree sufficiently to share a conception of justice. The success of this new positivism depends on how public the public culture in fact is: how much all reasonable comprehensive doctrines share the reasons that are said to be public.

Bibliography

Adams, Robert M. “The Theological Ethics of the Young John Rawls and Its Background.” In Rawls 2009, 24-101.

Adcock, Robert, and Mark Bevir. 2007. “The Remaking of Political Theory.” In Adcock, Robert, Mark Bevir and Shannon Stimson, eds. *Modern Political Science: Anglo-American Exchanges since 1880,* 209-33. Princeton: Princeton University Press.

Ambrose, Alice. 1935a. “Finitism in Mathematics I.” *Mind* 44: 186-203.

———. 1935b. “Finitism in Mathematics II.” *Mind* 44: 317-340.

Austin, John L. 1961. *Philosophical Papers*. Oxford: Clarendon Press.

———. 1962a. *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

———. 1962b. *Sense and Sensibilia*. Edited by G.J. Warnock. Oxford: Oxford University Press.

Axelrod, Robert. 1986. “An Evolutionary Approach to Norms.” *The American Political Science Review* 80: 1095-1111.

Aybar S., Harlan J. and Lee W. 1991. “John Rawls: For the Record.” *The Harvard Review of Philosophy* (Spring): 39-40.

Ayer, Alfred Jules. 1946. *Language, Truth, and Logic*. New York: Dover.

Baier, Kurt. 1953a. “Good Reasons.” *Philosophical Studies* 4: 1-15.

———. 1953b. “Proving a Moral Judgment.” *Philosophical Studies* 4: 33-44.

———. 1958. *The Moral Point of View*. Ithaca, NY: Cornell University Press.

Baillie, John, ed. 2002. *Natural Theology*. Eugene, OR: Wipf and Stock Publishers.

Berlin, Isaiah. 1973. “Austin and the Early Beginnings of Oxford Philosophy.” In Isaiah Berlin et al., eds. *Essays on J.L.* Austin, 1-16. Oxford: Clarendon Press.

Bevir, Mark. 1999. *Logic of the History of Ideas*. Cambridge: Cambridge University Press.

———. 2006. “Political Studies as Narrative and Science, 1880-2000.” *Political Studies* 54: 583-606.

Blumberg, Albert E. and Herbert Feigl. 1931. “Logical Positivism.” *The Journal of Philosophy* 28: 281-296.

Braithwaite, Richard Bevan. 1955. *Theory of Games as a Tool for the Moral Philosopher*. Cambridge: Cambridge University Press.

Broad, Charlie Dunbar. 1930. *Five Types of Ethical Theory*. New York: Harcourt, Brace & Co.

Brown, William Adams. 1902. *The Essence of Christianity*. New York: Charles Scribner’s Sons.
———. 1919 [1906]. *Christian Theology in Outline*. New York: Charles Scribner’s Sons.

Brunner, Emil. 1929. *The Theology of Crisis*. New York: Charles Scribner’s Sons.

———. 1934. *The Mediator: A Study of the Central Doctrine of the Christian Faith*. Translated by Olive Wyon. London: The Lutterworth Press.

———. 1939. *Man in Revolt: A Christian Anthropology*. Translated by Olive Wyon. New York: Charles Scribbner’s Sons.

Butler, Samuel. 1880. *Erewhon, or, Over the Range*. London: Ballantyne Press.

Campbell, Charles A. 1935. “Moral and Nonmoral Values.” *Mind* 44: 273-299.

Carnap, Rudolf. 1934. “On the Character of Philosophic Problems.” *Philosophy of Science* 1: 5-19.

———. 1935. *Philosophy and Logical Syntax*. London: Kegan Paul, Trench, Trubner & Co.

———. 1937. *The Logical Syntax of Language*. London: Kegan Paul.

———. 1942. *Introduction to Semantics and Formalization of Logic*. Cambridge, MA: Harvard University Press.

———. 1945a. “On Inductive Logic.” *Philosophy of Science* 12: 72-97.

———. 1945b. “The Two Concepts of Probability.” *Philosophy and Phenomenological Research* 5: 513-532.

———. 1959. “The Elimination of Metaphysics Through Logical Analysis.” Translated by Arthur Pap. In Alfred Jules Ayer, *Logical Positivism*, 60-81. Westport, CT: Greenwood Press.

———. 1969. *The Logical Structure of the World: Pseudoproblems in Philosophy*. Translated by Rolf A. George. Berkeley, CA: University of California Press.

———. 1987 [1932]. “On Protocol Sentences.” Translated by Richard Creath and Richard Nollan. *Noûs* 21: 457-70.

Craig, Leon H. 1975. “Contra Contract: A Brief against John Rawls’ *Theory of Justice*.” *Canadian Journal of Political Science* 8: 63-81.

Creath, Richard. 1987. “Some Remarks on ‘Protocol Sentences’.” *Noûs* 21: 471-75.

Delaney, Cornelius. 2003. “Realism, Naturalism, and Pragmatism.” In Thomas Baldwin, ed. *The Cambridge History of Philosophy 1870-1945*, 449-460. Cambridge: Cambridge University Press.

Dorrien, Gary. 2000. *The Barthian Revolt in Modern Theology: Theology without Weapons*. Louisville, KY: Westminster John Knox Press.

———. 2001. *The Making of American Liberal Theology: Imagining Progressive Religion 1805 – 1900*. Louisville, KY: Westminster John Knox Press.

———. 2003. *The Making of American Liberal Theology: Idealism, Realism, and Modernity 1900-1950.* Louisville, KY: Westminster John Knox Press.

Ducasse, Curt J. 1940. “The Nature and Function of Theory in Ethics.” *Ethics* 51: 22-37.

———. 1941. *Philosophy as a Science: Its Matter and Method*. New York: O. Piest.

Dummett, Michael. 1978. “Oxford Philosophy.” In Michael Dummett. *Truth and Other Enigmas*, 431-436. Cambridge, MA: Harvard University Press.

———. 1993. *Origins of Analytic Philosophy*. Cambridge, MA: Harvard University Press.
Dworkin, Ronald. 1973. “The Original Position.” *The University of Chicago Law Review* 40: 500-533.

Feigl, Herbert. 1954. “Method Without Metaphysical Presuppositions.” *Philosophical Studies* 5: 17-29.

Findlay, John Niemeyer. 1954. “The Justification of Attitudes.” *Mind* 63: 145-161.

Forguson, Lynd. 2001. “Oxford and the ‘Epidemic’ of Ordinary Language Philosophy.” *The Monist* 84: 325-345.

Frank, Phillip. 1946. *Foundations of Physics*. International Encyclopedia of Unified Science I:7. Chicago: University of Chicago Press.

Frankena, William. 1951. “Main Trends in Recent Philosophy: Moral Philosophy at Mid-Century.” *The Philosophical Review* 60: 44-55.

———. 1958. “Obligation and Motivation in Recent Moral Philosophy.” In A.I. Melden, ed. *Essays in Moral Philosophy*, 40-81. Seattle: University of Washington Press.

———. 1964. “Ethical Theory.” In R. Schlatter, ed. *Humanistic Scholarship in America: Philosophy*, 345-463. Englewood Cliffs, NJ: Prentice-Hall.

Freeman, Samuel. 2007a. *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*. Oxford: Oxford University Press.

———. 2007b. *Rawls*. New York: Routledge.

Galison, Peter. 1990. “Aufbau/Bauhaus: Logical Positivism and Architectural Modernism.” *Critical Inquiry* 16: 709-52.

———. 1997. “Constructing Modernism: The Cultural Location of the *Aufbau*.” In *Origins of Logical Empiricism*, edited by Ronald N. Geire and Alan W. Richardson, 17-44. Minneapolis: University of Minnesota Press.

Gauthier, David. 1987. *Morals by Agreement*. Oxford: Clarendon Press.

Geach, Peter. 1956. “Good and Evil.” *Analysis* 17: 32-42.

Gilligan, Carol. 1982. *In A Different Voice*. Cambridge, MA: Harvard University Press.

Greene, Theodore M. 1946. “Christianity and Its Secular Alternatives.” In Paul J. Tillich et al., eds., *The Christian Answer*, 72-127. London: Nisbet & Co.

Gregory, Eric. 2007. “Before the Original Position: the Neo-Orthodox Theology of the Young John Rawls.” *Journal of Religious Ethics* 35: 179-206.

Gutmann, Amy and Dennis Thompson. 1996. *Democracy and Disagreement*. Cambridge, MA: Harvard University Press.

Hacker, P.M.S. 1996. *Wittgenstein’s Place in Twentieth Century Analytic Philosophy*. Oxford: Blackwell Publishers.

Hagerstrom, Axel. 1953. *Inquiries into the Nature of Law and Morals*. Edited by Karl Olivecrona. Stockholm: Almqvist & Wiksell.

Hall, Thor. 1978. *Anders Nygren*. Waco, TX: Word Books Publishers.

Hampshire, Stuart. 1948. “Logical Necessity.” *Philosophy* 23: 339.

———. 1949. “Fallacies in Moral Philosophy.” *Mind* 58: 466-482.

Hanson, Norwood Russell. 1958. *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.

Hare, Richard M. 1952. *The Language of Morals*. Oxford: Clarendon Press.

———. 1963. *Freedom and Reason*. Oxford: Clarendon Press.

Harnack, Adolf von. 1901. *What is Christianity?* Translated by Thomas Bailey Saunders. Oxford: Williams and Norgate.

———. 1957. *Outlines of the History of Dogma*. Translated by Edwin Knox Mitchell. Boston: Starr King Press.

Hart, Herbert L.A. 1948-49. "The Ascription of Responsibility and Rights." *Proceedings of the Aristotelian Society* 49: 171-194.

Harrison, Jonathan. 1954. "When is a Principle a Moral Principle?" *Proceedings of the Aristotelian Society* Supp. Vol. 28: 111-34.

Hempel, Carl G. 1935. "On the Logical Positivits' Theory of Truth." *Analysis* 2: 49-59.

———. 1943. "A Purely Syntactical Definition of Confirmation." *Journal of Symbolic Logic* 8: 122-43.

———. 1945. "Studies in the Logic of Confirmation." *Mind* 54: 1-26.

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

Howard, Thomas Albert. 2000. *Religion and the Rise of Historicism*. Cambridge: Cambridge University Press.

Kiser, Edgar and Michael Hechter. 1998. "The Debate on Historical Sociology: Rational Choice Theory and Its Critics." *American Journal of Sociology* 104: 785-816.

Koikkalainen, Petri. 2009. "Peter Laslett and the Contested Concept of Political Philosophy." *History of Political Thought* 30: 336-359.

Korsgaard, Christine. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Laborde, Cécile, ed. 2002. "Rawls in Europe." *European Journal of Political Theory* 1, Special Issue.

Lamarque, Peter. 2010. "Wittgenstein, Literature, and the Idea of a Practice." *British Journal of Aesthetics* 50: 375-88.

Larmore, Charles. 2008. *The Autonomy of Morality*. Cambridge: Cambridge University Press.

Laslett, Peter and W. Garry Runciman, eds. 1963 [1956]. *Philosophy, Politics and Society*. Oxford: Basil Blackwell.

———. 1964 [1962]. *Philosophy, Politics and Society*, 2nd series. Oxford: Basil Blackwell.

———. 1967. *Philosophy, Politics and Society*. 3rd series. New York: Barnes and Noble.

Laslett, Peter and James Fishkin, eds. 1979. *Philosophy, Politics and Society*, Fifth series. New Haven: Yale University Press.

Malcolm, Norman. 2001. *Ludwig Wittgenstein: A Memoir*. New York: Oxford University Press.

Moore, George E. 1968 [1903]. *Principia Ethica*. Cambridge: Cambridge University Press.

Morris, Charles. 1946. *Signs, Language and Behavior*. New York: Prentice Hall.

Nagel, Ernest. 1936a. "Impressions and Appraisals of Analytic Philosophy in Europe. I" *The Journal of Philosophy* 33: 5-24

———. 1936b. "Impressions and Appraisals of Analytic Philosophy in Europe. II" *The Journal of Philosophy* 33: 29-53.

Neurath, Otto. 1944. *Foundations of the Social Sciences*. International Encyclopedia of Unified Science II:1. Chicago: University of Chicago Press.

Niebuhr, Reinhold. 1964 [1941]. *The Nature and Destiny of Man. Volume I: Human Nature*. New York: Charles Scribner's Sons.

Nielsen, Kai. 1959. "The 'Good Reasons Approach' and 'Ontological Justifications' of Morality." *The Philosophical Quarterly* 9: 116-130.

Nygren, Anders. 1953 [1932-1939]. *Agape and Eros*. Translated by Philip S. Watson. Philadelphia: The Westminster Press.

———. 1960. *The Essence of Christianity: Two Essays*. Translated by Philip S. Watson. London: The Epworth Press.

O'Neill, Onora. 1996. *Toward Justice and Virtue: A Constructive Account of Practical Reasoning*. Cambridge: Cambridge University Press.

Pogge, Thomas. 2007. *John Rawls: His Life and Theory of Justice*. Translated by Michelle Kosch. Oxford: Oxford University Press.

Polanyi, Michael. 1946. *Science, Faith, and Society*. London: Geoffrey Cumberlege.

Popper, Karl. 1959 [1934]. *The Logic of Scientific Discovery*. New York: Basic Books.

Princeton University. 1950. Princeton University Course Catalog, 1950-51. Princeton: Princeton University Press.

Putnam, Hilary. 1990. *Realism with a Human Face*. Edited by James Conant. Cambridge, MA: Harvard University Press.

Quine, Willard V.O. 1951. "Two Dogmas of Empiricism." *The Philosophical Review* 60: 20-43.

———. 1960. *Word and Object*. Cambridge, MA: M.I.T. Press.

———. 1969. "Epistemology Naturalized." In W.V.O. Quine, *Ontological Relativity and Other Essays*, 69-90. New York: Columbia Press.

———. 1981. "On the Nature of Moral Values." In W.V.O. Quine, *Theories and Things*, 55-66. Cambridge, MA: Harvard University Press.

———. 1986. "Reply to Morton White." In Lewis Hahn and Paul A. Shilpp, eds., *The Philosophy of W.V. Quine*, 664-665. La Salle, IL: Open Court.

———. 1990. "Homage to Rudolf Carnap." In Richard Creath, ed. *Dear Carnap, Dear Van*, 463-466. Berkeley, CA: University of California Press.

———. 1991. "Two Dogmas in Retrospect." *Canadian Journal of Philosophy* 21: 265-274.

Ramsey, Paul. 1950. *Basic Christian Ethics*. New York: Charles Scribner's Sons.

Rawls, John. 1946. "A Brief Inquiry into the Nature of Ethical Theory." John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 7, Folder 3.

———. c1947. "Remarks on Ethics" (Cover: J.Rawls Ethics). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 15.

———. 1950. "A Study in The Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral Worth of Character", Ph.D. diss., Princeton University.

———. c1950. "Delimitation of the Problem of Justice." John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 9 Folder 13.

———. [1950-2?]. "Ethics and Its Reasoning 1950-52 (?)."John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 4.

———. 1951. "Review." *The Philosophical Review* 60: 572-80.

———. c1952. “On Values.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7 Folder 9.

———. [1952?]. “Diseases of Ethical Reasoning.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 14.

———. [1952]. “Theory of Goods.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 2.

———. 1952-3. “On Explication Oxford 1952-3.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 18.

———. 1953a. “Oxford Notes, Spring 1953.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 10.

———. 1953b. “Justice as Fairness, Cornell Seminar 1953 Fall.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 7, Folder 11.

———. [1953?]a. “Wittgenstein investigation, lexicon.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 60.

———. [1953?]b. “Wittgenstein Criteria.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 8.

———. [1953?]c. “Wittgenstein Investigations.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 2.

———. 1954a. “Christian Ethics: Class at Cornell.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 5.

———. 1955. “Two Concepts of Rules.” *The Philosophical Review* 64: 3-32.

———. 1956. “Rational Choice and the Concept of Goodness.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 9, Folder 3.

———. 1958a. “Moral Feeling I (1958).” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 34, Folder 19.

———. [1958-1962]. “Essay V.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 1.

———. [1958]. “Moral Judgment, Relativism.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 8, Folder 8.

———. 1960. “Political Philosophy 171.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35 Folder 12.

———. 1960. “Moral Feelings, 1960.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 1.

———. [1965]. “Goodness as Rationality.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 17.

———. 1962. John Rawls, “Political Philosophy 171, 1962.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folders 8-13.

———. 1964. “Moral Psychology, 1964-5.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 6.

———. 1964b. “Essay on Justice. First Draft of *A Theory of Justice*, 1 of 2.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 17, Folder 2.

———. 1965. “Justice, second draft of *A Theory of Justice*. March 1965.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 17, Folder 4.

———. 1965. “Philosophy 171. Chapters on Justice. Draft of *A Theory of Justice* reproduced to students” (1965 Fall). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 18, Folder 4.

———. 1965. “Natural Law, 1962, 1965” (1965). John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 35, Folder 13.

———. 1966a. “Philosophy 171, Lectures I-IV 1966-1967.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 36, Folder 10.

———. 1966-7a. “Analytic Ethics and Justification. 1966-7.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 5, Folder 6.

———. 1967. John Rawls, “Justice as Fairness II”. John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 10, Folders 1-5.

———. 1970. “Philosophy 169. Lectures I-IV.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 5, Folder 2.

———. 1970. “Philosophy 169. Part II. Lectures V-IX.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 5, Folder 3.

———. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

———. [1972]. “Reply to Stanley Moore.” John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 19, Folder 4.

———. 1978. “Letter to Robert Audi.” John Rawls Faculty Papers, Harvard University Archives, HUM 48, Box 43, Folder 1.

———. [1993?]. “Autobiographical Notes.” John Rawls Faculty Papers, Harvard University Archives, HUM 48 Box 42, Folder 12.

———. 1999a. *Collected Papers*. Edited by Samuel Freeman. Cambridge, MA: Harvard University Press.

———. 1999b. *A Theory of Justice* (revised ed.). Cambridge, MA: Harvard University Press.

———. 2001a. *The Law of Peoples*. Cambridge, MA: Harvard University Press.

———. 2001b. *Justice as Fairness: A Restatement.* Cambridge, MA: Harvard University Press.

———. 2001c. “Afterword: A Reminiscence.” In Juliet Floyd and Sanford Shieh, eds. *Future Pasts: The Analytic Tradition in Twentieth Century Philosophy*, 417-430. Oxford: Oxford University Press.

———. 2005 [1993]. *Political Liberalism*. New York: Columbia University Press.

———. 2009. *A Brief Inquiry into the Meaning of Sin and Faith*. Edited by Thomas Nagel. Cambridge, MA: Harvard University Press.

Reichenbach, Hans. 1938. *Experience and Prediction*. Chicago: The University of Chicago Press.

Reidy, David. 2010. “Rawls’s Religion and Justice as Fairness.” *History of Political Thought* 31: 309-343.

Richardson, Alan W. 2003. “Logical Empiricism, American Pragmatism, and the Fate of Scientific Philosophy in North America.” In Gary L. Hardcastle and Alan W. Richardson, eds. *Logical Empiricism in North America*, 1-24. Minneapolis, MN: University of Minnesota Press.

Ritschl, Albrecht. 1902. *The Christian Doctrine of Justification and Reconciliation*. Translated by H.R. MacIntosh and A.B. Macaulay. Edinburgh: T. & T. Clark.

Rorty, Richard, ed. 1992 [1967]. *The Linguistic Turn: Essays in Philosophical Essays*. The University of Chicago Press: Chicago.

———. 2005. “How Many Grains Make a Heap?” *London Review of Books* 27:2.

Rousseau, Jean-Jacques. 1979. *The Social Contract.* Translated by Maurice Cranston. London: Penguin.

Ross, William D. 1930. *The Right and the Good*. Oxford: Clarendon Press.

Rueff, Jacques. 1929. *From the Physical to the Social Sciences: Introduction to a Study of Economic and Ethical Theory*. Translated by Herman Green. Baltimore: John Hopkins University Press.

Rumscheidt, H. Martin, ed. 1972. *Revelation and Theology: An analysis of Barth-Harnack Correspondence of 1923*. Cambridge: Cambridge University Press.

Ryle, Gilbert. 1971. "Philosophical Arguments." In Gilbert Ryle, *Collected Papers*, vol. II, 194-211. London: Hutchison & Co.

Sandel, Michael J. 1982. *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.

Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scheffler, Samuel. 1979. "Moral Independence and the Original Position." *Philosophical Studies* 35: 397-403.

Schleiermacher, Friedrich. 1996. *On Religion: Speeches to its Cultured Despisers*. Translated and edited by Richard Crouter. Cambridge, Cambridge University Press.

Schlick, Moritz. 1939. *Problem of Ethics*. Translated by David Rynin. New York: Prentice Hall.

———. 1959. "The Turning Point in Philosophy." Translated by David Rynin. In Alfred Jules Ayer, ed. *Logical Positivism*, 53-59. Westport, CT: Greenwood Press.

Skorupski, John. 1990. "The Legacy of Modernism," *Proceedings of the Aristotelian Society* 91: 1-19.

Sluga, Hans. 1998. "What Has History Got to Do with Me? Wittgenstein and Analytic Philosophy." *Inquiry* 41: 99-121.

Soames, Scott. 2003. *Philosophical Analysis in the Twentieth Century*, vol. II. Princeton: Princeton University Press.

Stace, Walter Terrence. 1939. *The Concept of Morals*. New York: the Macmillan Company.

Stevenson, Charles L. 1944. *Ethics and Language*. New Haven, CT: Yale University Press.

Strawson, Peter. 1959. *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.

Taylor, Robert S. *Reconstructing Rawls: the Kantian Foundations of Justice as Fairness*. University Park, PA: the Pennsylvania State University Press.

Thomas, George F. 1946. “Central Christian Affirmations.” In Paul J. Tillich et al., *The Christian Answer*, 128-180. London: Nisbet & Co.

Toulmin, Stephen. 1950. *An Examination of the Place of Reason in Ethics*. Cambridge: Cambridge University Press.

Tully, James, ed. 1998. *Meaning and Context: Quentin Skinner and his Critics*. Princeton: Princeton University Press.

Uebel, Thomas E. 1996. "Anti-Foundationalism and the Vienna Circle's Revolution in Philosophy." *British Journal for the Philosophy of Science* 47: 415-440.

Urmson, James Opie. 1950. “On Grading.” *Mind* 59: 145-169.

von Neumann, John and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Walzer, Michael. 1983. *Spheres of Justice: A Defense of Pluralism and Equality*. New York: Basic Books.

Williams, Bernard. 2003. “The Spell of Linguistic Philosophy.” Interview with Bryan Magee. Princeton, NJ: Films for the Humanities & Sciences.

———. 2011 [1972]. *Morality: An Introduction to Ethics*. Cambridge: Cambridge University Press.

Wisdom, John. 1936. “Philosophical Perplexity.” *Proceedings of the Aristotelian Society* 37: 71-88.

Wittgenstein, Ludwig. 1958 [1953]. *Philosophical Investigations*. Translated by Gertrude E.M. Anscombe. Upper Saddle River, NJ: Prentice Hall.

———. 1965. *The Blue and Brown Books/ Preliminary Studies for the “Philosophical Investigations”*. New York: Harper Torchbooks.

———. 1969. *On Certainty*. Edited by G.E.M. Anscombe and G.H. von Wright and translated by Denis Paul and G.E.M. Anscombe. New York: Harper Torchbooks.

Wolff, Robert Paul. 1977. *Understanding Rawls: A Reconstruction and Critique of A Theory of Justice.* Princeton: Princeton University Press.