Part I

Ambitions

1

From Philosophical Theology to Democratic Theory

Early Postcards from an Intellectual Journey

DAVID A. REIDY

It is easy to kill a subject by demanding too much of it early on; a subject needs to be guided by big intuitive ideas, particularly at the start. . . . It is a delusion to think that rigorous analysis in a small area unguided by a large idea is of much value. One does not understand even a small thing in this way.
John Rawls, 1964, to students in his moral philosophy course

1. Introduction

Rawls published *A Theory of Justice* in 1971, though, as he noted in the "Preface," he had been circulating and teaching from earlier drafts through much of the 1960s. But *TJ* does not really originate in the 1960s. Its roots run at least to the late 1940s and 1950s, a period covering Rawls's years as a graduate student and then lecturer at Princeton, his year (1952–1953) as a visiting fellow at Oxford during which he first saw clearly the project that would occupy him for some 50 years, and his time as a faculty member at Cornell. Indeed, in some respects, its roots run to the late 1930s and early 1940s, the time of Rawls's undergraduate study at Princeton and his work on his now published undergraduate thesis (*BI*). While it was not until the mid-1960s that Rawls had in hand all the essential elements of the "painting" he sought to share in *TJ* and to complete in later works, he was possessed of, and by, the core of the "vision" by the mid to late 1950s.[1] And he was in important ways oriented toward it even earlier than that. In several respects, then, *TJ* is an early mid-century book.

This is a fact naturally overlooked by readers. *A Theory of Justice* did not reach a general and wide philosophical audience until the 1970s, a time very different philosophically, politically and culturally from that of its origin. While *TJ* clearly addresses some concerns central to that time, for example, civil disobedience and conscientious objection, in general it is not

A Companion to Rawls, First Edition. Edited by Jon Mandle and David A. Reidy.

profitably read primarily against the concerns, expectations and cultural landscape of the 1970s or even the latter 1960s.

In the early mid-century many thoughtful Americans were anxious about the viability of the sort of inclusive, mass democracy that seemed to be taking root. Further, many American liberals, troubled by the ideologically motivated disasters of World War II and the Soviet system, were both eager to distance themselves from political self-understandings grounded in or animated by big ideas and inclined favorably toward more modest political self-understandings of liberal democracy as either a regulated or civilized struggle among competing interest groups or a mechanism for the rational aggregation of private preferences with an eye toward efficiency (see, e.g., Schumpeter 1942; Downs 1957).

Rawls too was anxious about the viability of the sort of inclusive mass democracy that seemed to be taking root. He worried that the country lacked the resources necessary to sustain the requisite public trust among citizens. And Rawls too was deeply influenced by the ideologically driven disasters of World War II and the Soviet system. But he worried that the modest political self-understandings on offer to Americans were not only insufficient to sustain the requisite public trust among them but also in an enduring way to draw their stable allegiance as free equals. He further worried that if internalized and effectively regulative, the political self-understandings on offer would have a corrosive effect on persons, hindering rather than helping them to realize themselves as persons in community. He sought for a polity of free and equal citizens a political self-understanding animated by a big intuitive idea capable of underwriting genuine public trust among them, reliably drawing their enduring allegiance and contributing to their self-realization as persons in community. His goal was a big intuitive idea with universal reach. *A Theory of Justice* is a giant first step toward expressing this idea.

There are aspects of this idea that account, I think, for some of its gravitational force, as it were – its capacity to draw persons into its gravitational field – that can be traced back to Rawls's very early work as an undergraduate in philosophical theology. And there are also aspects that can be traced back to his early work as a graduate student in moral philosophy. In this essay, I take up Rawls's journey from philosophical theology through moral philosophy to democratic theory and political philosophy and pause at, to reflect on, a few significant points early in the journey. My aim is to give a sense – I can offer here no more than what amounts to postcards – of some of Rawls's important early concerns and commitments that structure or at least cast significant shadows over his later work in political philosophy, *A Theory of Justice* and subsequent works.

I do not mean to suggest that Rawls's journey to *TJ* is marked by or best understood in terms of only the concerns and commitments I discuss. There are others. For example, Wittgenstein was a very important influence on Rawls for many decades – Rawls acknowledges this in several letters. I do not discuss this influence here. Nor do I mean to deny important discontinuities in the development of Rawls's thought over time. His movement in the 1950s away from both theism and from a kind of Millian utilitarianism merit mention here, as does his introduction in the 1980s of the family of ideas associated with political liberalism, though in this latter case I think there is less discontinuity than is often alleged. I mean only to suggest that *A Theory of Justice* is a more rewarding read if one attends to the aspects of Rawls's big intuitive idea that are set out here and if one keeps in mind their origins in his early-mid-twentieth century thinking.

2. The Philosophical Theology of the Undergraduate Thesis

As is now well-known, Rawls's Princeton undergraduate philosophy thesis, "A Brief Inquiry into the Meaning of Sin and Faith: An Interpretation Based on the Concept of Community," sits at the intersection of philosophical theology and theological ethics.[2] It argues against familiar understandings of sin and faith rooted in a conception of God as our highest good and so the proper and ultimate object of our rational desire. It argues in favor of understandings of sin and faith rooted in a conception of God as the complete, self-sufficient and eternal instantiation of personality and community, neither of which can exist without the other, as well as in our capacities for participation in personality and community, dependent for their realization on the unconditional grace of God though they may be. But it is not primarily this theological content to which I wish to draw attention here, though I do want to draw attention to Rawls's conception of personality and community. Mainly it is the methodological and meta-theological context within which Rawls works out his view that I think merits notice.

Rawls's view draws on and synthesizes aspects of two traditions in philosophical theology and theological ethics with wide currency in the early twentieth century. The first is biblical historicism. The second is the neo-orthodoxy sometimes associated with what came to be known as "theology of crisis." Biblical historicism arises in the late eighteenth and early nineteenth centuries in response to the fact that it was increasingly difficult to deny that the Bible was written by many people over an extended period of time in various contexts. Informed by these facts, those inquiring into the "meaning" of the Bible found it increasingly difficult to represent it as the articulation of a single, focal revelation that took place at one time and through one person. Biblical historicism responds to this difficulty by taking the Bible as the record of Christian experience over time and seeking through reasoned analysis of that experience, as lived in historical context, a universal truth underlying and unifying it over time and so expressing the "meaning" of the Bible or the enduring universal truth of Christianity. The Hegelian idea that Christianity expresses finally and completely the universal rational meaning of moral and religious experience is kin to nineteenth-century biblical historicism.

Neo-orthodoxy or "theology of crisis" emerges as a reaction both from within and against biblical historicism in the early twentieth century, and in particular after World War I. Skeptical of biblical historicism's ability reliably to deliver through its methods the enduring universal truth of Christianity and cognizant of the tremendous harm humans are capable of if not guided by that truth, the proponents of neo-orthodoxy put the emphasis back on the authority of the Bible as the record of a unique – one time, one person – revelation, the content of which was not accessible by human reason alone. Interestingly, some associated with neo-orthodoxy proposed an account of the content of this revelation that Rawls ultimately found quite congenial, namely that God is the complete, self-sufficient and enduring realization of personality and community, each of which requires the other; that all personal relations begin with an opening by one person to another, an invitation to community; that Christ is that invitation to communion with God; and that the invitation and our capacity to overcome pride and accept it are both functions of God's grace.

In his undergraduate thesis, Rawls appropriates this neo-orthodox account of the content of the revelation given in the Bible and argues, within a biblical historicist framework, that

it best accounts for – as a kind of deep explanation of – not only Christian experience, including his own, but also the experiences of non-Christians. He pays special attention to the experiences of conversion through grace and to the forms of aloneness and despair experienced, and capacities for harm and evil exhibited, by those who have not converted. Conversion here refers to the process, itself a gift of grace, of being saved from the deep tendency in our nature toward prideful refusal of personal relations and community, of being brought to a genuine openness and orientation to both with others, including with God. Here conversion involves no affirmation of doctrinal theological content or dogma but rather a reorientation of one's moral psychology and self-understanding and so one's experience of living with others in the world. Reason and linguistic communication more generally serve as necessary media or instruments of personal relations and community, but they do not initiate or demand either. The call to personal relations and community originates elsewhere.

There are several things to notice here. One is the fact that Rawls seeks the meaning of Christianity, of sin and faith, in the best explanation – not causal explanation, but rational explanation – of Christian experience, indeed of human experience generally. Undoubtedly, he takes the Bible as an authoritative expression or record of Christian (as well as Hebrew and other) experience. But it is the experience, not the biblical text per se, that constitutes the data to be understood. And the experience is to be understood not in the space of causes, as it were, but in the space of a rational, in the sense of intelligible to a common or shared reason, moral psychology. Indeed, though Rawls does not say so explicitly, it is not unreasonable to suppose that he understood the articulation of such a rational explanation as itself a practical contribution to the realization of the universal Christian community. Without it, we cannot fully understand ourselves and so cannot with full awareness or understanding participate in communion with one another and with God.

Another thing to notice here is that notwithstanding the exclamation point that the early stages of World War II placed behind the "crisis" to which the neo-orthodox "theology of crisis" turn to the revealed authority of scripture was a response, Rawls is not drawn in his undergraduate thesis to this aspect of neo-orthodoxy. He makes no appeal to what he will a decade later in his PhD dissertation refer to as "exalted authorities." The Bible is authoritative, but it is so as a record of a certain pattern of human experience. And it is this pattern of experience that must be rendered intelligible and properly understood in the space of reasons. If there is any appeal to authority here, it is the authority of our own self-recognition and self-understanding. Effectively Rawls asks us in his undergraduate thesis whether, having heard his explanation of the experiences of Christians and others and put it side by side with the story he attributes to Augustine and Aquinas, we do not in fact more fully and completely recognize and understand ourselves in and through his story than its alternative. There is no suggestion that this self-recognition and self-understanding by itself will or can bring about the experience of "conversion" at the center of Christian experience, an experience that is itself, on Rawls's account, a gift of grace. Nevertheless, it is essential to participating with full understanding in community with others and with God.

3. Ethics as Science

After returning from his service in World War II, Rawls began graduate study in philosophy at Princeton. In 1946, as a first year graduate student, he wrote "A Brief Inquiry into the

Nature and Function of Ethical Theory." He begins by asking what it is that moral philosophers do. He argues that the way to answer the question is not to survey, but to observe moral philosophers. If we observe them, we find that they are engaged in a science of moral judgment. They seek to explain competent moral judgments in a way that would enable us reliably to predict them. They do not seek the meaning of moral terms, in the sense of identifying synonyms which might be substituted for them in any statement in which they appear without altering its truth value. Nor do they seek to uncover what one intends to assert or has in one's mind when one makes a moral judgment. Nor, further, do they seek to identify the logically basic objects and relations ingredient in the propositions expressed by moral judgments, or to select among rival logical or formal notations we might use in talking about those objects and relations. In short, Rawls concludes, moral philosophers do none of the things that had come to be associated mid-century with the tradition of analytic philosophy associated with Bertrand Russell, G.E. Moore and the early Wittgenstein. Instead, what moral philosophers do is to construct theoretical models to explain and predict familiar, everyday, noncontroversially competent moral judgments. Rawls refers to this work as "explication" and "ethics as science."

Of course, Rawls notes, this is work one might think more profitably done by psychologists, sociologists or anthropologists (or, today, by socio-biologists or neurologists). But moral philosophers do not approach moral judgments as mere events, mental or otherwise, to be explained and predicted within the natural causal order. Rather, they approach them as, well, *moral judgments* – as the publicly visible manifestation of the exercise or activity of practical reason by and among persons. Accordingly, Rawls maintains, the kinds of models of interest to the moral philosopher are something like "reasoning machines" – systems of definitions and axioms such that when fed determinate input regarding the sorts of familiar moral choices with respect to which we can noncontroversially distinguish competent from incompetent judgments, they yield theorems, or moral principles, that provide sufficient reasons for, and thereby render intelligible to us, all and only the competent judgments. These "reasoning machines" are not meant to represent or to be incorporated necessarily into any actual psychological process, what actually goes on in the mind of a person making a moral judgment. Rawls notes that very often we make competent moral judgments without, or without any awareness of, any deliberative thought process at all. We simply hear or express the "voice of conscience." This the moral philosopher regards as, in itself, no defect in need of correction. Instead, the moral philosopher aims to *represent* the phenomenon of competent moral judgment as a public or visible manifestation of the activity or exercise of practical reason. The moral philosopher offers us a way of understanding ourselves, of making ourselves intelligible to one another, as persons, as rational social beings with a capacity for moral judgment. He will later in his dissertation, but does not yet in this 1946 paper, take up the contribution such a self-understanding or representation makes, when not only shared but internalized by a group of persons, to the realization of personal relations in community, relations best characterized in terms of mutual justification.

Rawls does not claim that our ability to make or to identify familiar, everyday, noncontroversially competent moral judgments depends on our successfully explicating those judgments through appeal to a "reasoning machine" worked up within ethics as science. He gestures, already in 1946, to the linguistic analogy, noting that our ability to utter and identify grammatically sound sentences seems not to depend at all on our ability to represent that ability in terms of a system of grammar worked up within linguistics as science. What is at

stake in ethics as science is not our ability to make or even to identify familiar, everyday, noncontroversially competent moral judgments, but rather the nature, content and implications of our self-representation and self-understanding as beings exercising this ability.

To the extent that Rawls locates ethics as science in relation to the analytic tradition associated with Russell, Moore and the early Wittgenstein, he does so by reference to Frege's work in deductive logic, which Rawls regarded as a scientific explication, within the space of reasons rather than causes, of familiar, everyday, noncontroversially competent judgments regarding valid inference. Frege successfully represented our ability to make and identify valid inferences in terms of a "reasoning machine" capable of explaining and reliably predicting our judgments regarding valid inference. In so doing, he put us in a position to understand our noncontroversially competent judgments regarding valid inference as the public or observable manifestation of the activity or exercise of our capacity for theoretical reason. Rawls took his task to be doing for noncontroversially competent moral judgments what Frege did for noncontroversially competent judgments regarding valid inference.

Rawls identifies four main activities or exercises of reason, two theoretical and two practical. There is for each a "science" within the space of reasons that is properly pursued by philosophers as persons who seek to understand themselves as rational and social animals capable of knowledge and morality. Theoretical reason expresses itself not only in the form of noncontroversially competent judgments regarding valid inference, but also in the form of noncontroversially competent judgments regarding theory confirmation. Frege developed a "science" of the former: deductive logic; J.S. Mill advanced a "science" of the latter: inductive logic. Practical reason expresses itself not only in the form of noncontroversially competent judgments regarding one's ends and the means to them, what Rawls would later call "the rational," but also in the form of noncontroversially competent judgments regarding one's relation as a person with other persons in social life, or what he would later call "the reasonable." The philosophical "science" of the former is the theory of rational choice, understood to include axiology. The philosophical "science" of the latter Rawls sought to advance through ethics as science.

Rawls characterizes his concern with and ambition for ethics as science as fully scientific in the sense associated with the Vienna Circle. That ethics as science unfolds within the space of reasons rather than causes is irrelevant to its status as science. What is crucial, Rawls insists, is that it must avoid all theoretical claims that can be neither confirmed nor refuted by publicly observable evidence, in particular the evidence constituted by familiar, everyday, noncontroversially competent moral judgments. If there is to be a sound moral science, it must proceed in this fashion. Its aims are descriptive, explanatory and predictive, rather than normative or prescriptive. Or, better, a sound moral science is normative only in the nonprescriptive sense of providing a clear and intelligible rational model or representation of noncontroversially competent moral judgments as a publicly observable phenomenon. That is, it provides us with a predictively reliable way of representing – a criterial understanding of – the distinction between noncontroversially competent moral judgments and noncontroversially incompetent moral judgments.

Rawls insists that we ought to draw no metaphysical conclusions from any success of ethics as science. From the fact that an intelligible reasoning machine generates theorems that make it possible reliably to predict all and only noncontroversially competent moral judgments, no metaphysical conclusions follow. There is no reason to think that the reasoning machine or the theorems it generates tells us anything about what ultimately or "really"

makes such judgments noncontroversially competent, about the "essence" of morality or of moral rightness, about the so-called "right-making" properties of the world. Rawls dismisses inquiries into these matters as stemming from little more than the traditional "quasi-religious" character of moral language and judgment.

On the other hand, if ethics as science succeeds, then, as Rawls notes, emotivist and other noncognitivist orientations toward moral judgment thought to follow from a generally scientific philosophical orientation would be destroyed from within. So while Rawls insists there are no positive metaphysical conclusions to be drawn from the success of ethics as science, there are some negative metaphysical conclusions to be drawn. Certain possibilities are preserved. Others are ruled out. One of Rawls's earliest impulses, then, would appear to be to try to save a properly scientific philosophical orientation from various excesses to which it seemed all too easily tempted.

The shadow of the Vienna Circle and scientific philosophy more generally falls over Rawls's 1946 "Brief Inquiry" paper in another respect. Rawls takes the view that the "meaning" of any moral term is best understood as given by the explication – the "reasoning machine" and the theorems or moral principles it generates – that most reliably predicts as intelligible and justifiable all and only the noncontroversially competent moral judgments in which it figures. That is, the "meaning" of a moral term is given by the "scientific theory" – responsive only to public, observable evidence, albeit worked out in the space of reasons – that successfully accounts for the noncontroversially competent moral judgments in which it appears. It is not given by synonyms that might be substituted for it within any statement in which it appears without any change to that statement's truth value. Such inquiries into "synonymy meaning" provide no independent basis for predicting noncontroversially competent moral judgments and thus no independent basis for understanding the actual meaning of the moral terms that occur within them. The sense in which such inquiries into "synonymy meaning" shed light on the actual meaning of moral terms is shallow and of no more than linguistic interest. Nor is the actual meaning of moral terms given by some private mental content thought to be present to the mind when moral terms are competently used. The pursuit of meaning as private mental content is, Rawls insists, a hopeless pursuit.[3]

As a model for his ethics as science, Rawls invoked Hans Kelsen's "pure theory of law." Kelsen, the Austrian legal positivist, aimed at a "legal science" capable of laying out, within the space of reasons, the norms governing valid law and competent legal judgment, without moralistic assumptions or tendencies and without naturalistically reducing either phenomenon to only facts and causes. Kelsen held that as a normative social practice, the law could not be fully understood solely in terms of facts and causes, and thus could not be fully understood solely within or through the social sciences. As a normative social practice, it had to be understood also within the space of reasons, in terms of norms or principles of competent (legally valid) judgment. But to understand it within the space of reasons, one had to approach it scientifically and without moralistic assumptions or aspirations. Legal philosophy, for Kelsen, was just another name for the scientific study of law as a normative social phenomenon within the space of reasons (see Kelsen 1978 [1934]).

Rawls conceived of moral philosophy on analogous terms. Just as a "pure" theory of law would identify the most basic norms or reasons in terms of which we might, within any particular legal system, render intelligible and reliably predict noncontroversially competent legal judgments, so too a "pure" theory of ethics would identify the most basic norms or

reasons in terms of which we might render intelligible and reliably predict noncontroversially competent moral judgments. Just as what makes a legal judgment noncontroversially competent is that it is a judgment universally shared or nearly so among those engaged in the practice of law and is arrived at and remains stable under favorable background conditions (a stable legal system, the absence of bribes, etc.), so too what makes a moral judgment noncontroversially competent is that it is a judgment universally shared or nearly so among those engaged in the practice of morality and is arrived at and remains stable under favorable background conditions (a free society, without disfiguring social forces, within which material and other conditions essential to moral development are secure, etc.). The task of moral philosophy is to represent rationally the competent moral judgments free and intelligent persons naturally make, not to establish for them the competency of their moral judgments.

Rawls concluded his 1946 paper by suggesting in outline an ethical theory that seemed to him promising, at least "if ethics is to be done as a science." He called it "imperative utilitarianism." The theory purports to cover only noncontroversially competent judgments of right, and then only at the level of individual actions. It represents such judgments as the result of reasoning from principles established by a reasoning machine that marks as required or forbidden action types that persons are not likely to perform or refrain from performing without significant social incentives, and with respect to which, at least in their statistically most common occurrences, social utility strongly depends on their being, respectively, performed or not performed. Of course, as Rawls acknowledges, one must feed such a reasoning machine a great deal of information provided by the natural and social sciences in order to have any chance of generating principles that would provide sufficient reasons to support noncontroversially competent moral judgments of right in this domain. The machine must be fed information about individuals' beliefs, about the statistical patterns of their behavior (with and without incentives), about the utility produced by various action types and the costs associated with making instrumental use of morality as a social incentive to induce or deter such action types, and so on. And this information will be vast and vary from society to society. But, Rawls maintains, at least in 1946, this is no defect.[4] Indeed, it is a merit insofar as the theory may be applied to any society or population, no matter what the facts about its members' conduct, utility profiles, and so on. The only question is whether once fed the relevant information the reasoning machine generates principles that make it possible reliably to predict noncontroversially competent moral judgments of right for the relevant population. If it does so across all societies or populations, then it gives us the "meaning" of right, simply *qua* right, by giving us a "scientific" (answering to public observable evidence only) theory of competent moral judgments for any society.[5]

The largely theoretical, or "scientific," orientation and concerns of Rawls's 1946 paper is one of its striking features. The paper evidences no significant or sustained interest in ethical theory as a means of improving or correcting moral judgments or of perfecting moral capacities or dispositions more generally. And while it aims at the production, via a "reasoning machine," of moral principles capable of justifying – that is, functioning as intelligible public reasons for – competent moral judgments, and so of representing competent moral judgments as rational and cognitive, it does not address the question of the justification of the principles themselves or the "reasoning machine" from which they are derived. Nor does it take up the quality of the relations realized among persons who affirm one and the same scientific representation of their capacity for competent moral judgment. These matters will come to occupy Rawls's attention within just a few years in his PhD dissertation.

Perhaps anxious about the largely theoretical or "scientific" orientation of the paper, Rawls notes in its final pages that, as with all theoretical scientific inquiries, so too with ethics as science: "men of affairs" will want to know its "cash value." He cites two benefits. The first arises in response to the fact that we sometimes find ourselves faced with situations so novel that we have no clear sense of what would count as a competent moral judgment – the voices of both individual conscience and the community's moral tradition are silent. In these cases, we may find value in an appeal to the principles of an ethical theory otherwise reliable as a predictor of noncontroversially competent moral judgments, for they may speak where individual conscience and community standards are silent. They may help at least to orient deliberation and discussion.

The second arises in response to the fact that people sometimes disagree in their moral judgments. When they do, it does not follow necessarily that anyone is guilty of a less than competent exercise of her moral capacities or that anyone is doing something that may not be represented as rational and cognitive. Of course, these are nevertheless possibilities. Sometimes one or both parties to a moral disagreement venture an incompetent moral judgment, one that could not be represented as rational or cognitive. Moral disagreements often arise, on one or both sides, from naked class ambition, group bias, or simple old-fashioned selfishness swamping the competent exercise of moral capacities. But there is no reason to suppose that they may not sometimes arise reasonably for other reasons. So there are reasonable moral disagreements and what we might call simple moral disagreements. But it is not always easy to distinguish the former from the latter. Yet we have compelling reasons to want to be able reliably and publicly to do so. For those driven by naked class ambition, group bias, or old-fashioned selfishness to incompetent moral judgments and so to simple moral disagreements with others will often, perhaps typically, take great pains to pretend at being reasonable. Without the ability to distinguish reliably and publicly between reasonable and simple disagreements, a community is likely, then, to misunderstand the content of its own established community standards or respected voices of conscience. An ethical theory, properly worked up within ethics as science, ought, Rawls maintains, to "enable us to say to the disputants . . . what their moral judgments really are" and "to expose moral pretense . . . for what it really is, naked class ambition, group bias, or selfishness."

Rawls does not think of the "cash value" of ethics as science in terms of replacing established community standards or respected voices of conscience as a source of moral guidance. Nor does he think of it in terms of drawing moral agreement out of the soil of genuine or reasonable moral disagreement. Morally decent people sometimes reasonably disagree. Not all moral disagreements are really disagreements rooted in naked class ambition, group bias, or selfishness merely pretending to be genuine or reasonable disagreement. Genuine or reasonable disagreements are, it seems, problems for which ethics as science is not the solution, except insofar as ethics as science contributes to our ability to distinguish them from unreasonable disagreements that merely appear, or are unreasonably made to appear, to be reasonable.

To illustrate his claim regarding the "cash value" of ethics as science, Rawls cites the postwar conflict, already significant in 1946, over the civil rights of African-Americans. A sound ethical theory, he suggests, would enable us publicly to expose this dispute for what it really is: an unreasonable disagreement arising out of the unreasonable assertion of naked class-ambition or group-bias dressed up to have the mere appearance of competent moral judgment. Those opposed to recognizing the civil rights of African-Americans, often occupying positions of power and commanding social respect, may pretend to be exercising their

moral capacities. They may use moral language, talk in cool tones, and even invoke what may seem to be moral principles. But their pretense will not survive an encounter with the principles of a sound ethical theory. And it must be exposed for what it is if the community is to understand and to have confidence its own established standards and respected voices of conscience.

Rawls is confident in 1946 that a successful explication (along the lines of his favored "imperative utilitarianism") of the noncontroversially competent moral judgment of free and intelligent men and women living in more or less favorable social conditions will yield principles capable of supporting only the judgment that African-Americans must be granted full and equal civil rights. Contrary judgments may be encouraged by rhetoric, propaganda, and emotive appeals and may be voiced in a moral idiom or key. But they will not be vindicable, from the point of view of ethics as science, as competent exercises of our moral capacity understood as a rational cognitive capacity. If they pretend to this status, they are counterfeit. As a "reasonable" moral disagreement, the disagreement over the civil rights of African-Americans is counterfeit. Herein ethics as science shows its "cash value."

4. From Ethics as Science to Moral Philosophy

Rawls first moves beyond ethics as science to take up the question of the justification of the moral principles it establishes and the reasoning machine it puts to use in a late 1940s paper titled "Ethical Rationalism." But it is not until the 1950 dissertation that the move is largely completed and given more or less full expression. There he first indicates an interest in not only the representational capacity but also the regulative and motivational capacity, as part of an internalized self-understanding, of the moral principles and reasoning machine given by ethics as science. I focus here on the dissertation.

In his 1950 dissertation, "A Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments of the Moral Worth of Character," Rawls builds on the earlier papers discussed above and ventures a more thoroughly systematic treatment of the rationality of our moral judgments and capacities. He sets out an approach to moral philosophy which he then applies, by way of illustration, to judgments regarding the moral worth of character. He expresses his intention to extend the approach taken, in due course, to judgments regarding ends or goods (questions of value) and to judgments regarding right actions. He pays no attention to judgments regarding institutions or practices. Indeed, he seems to assume at this point that a moral theory is complete so long as it ranges over judgments of character, value and right action.[6]

Rawls distinguishes between two tasks essential to moral philosophy. The first is to explicate noncontroversially competent moral judgments in terms of general moral principles given by an intelligible reasoning machine. This is the now familiar task of ethics as science. Rawls acknowledges in his dissertation that ethics as science may yield more than one candidate explication of noncontroversially competent moral judgments. And so moral philosophy must move beyond ethics as science to include the systematic development and comparison of different candidate explications of noncontroversially competent moral judgments. This work Rawls dubs "ethical theory" or "moral theory." Ethical theory or moral theory starts with but takes us beyond ethics as science by helping us to refine and deepen our understanding of plausible candidate explications. Plausible candidate explications are

worked up more fully into "moral conceptions." The aim of ethical theory or moral theory, however, is not simply to work up a more complete understanding of plausible candidate explications. It is also to put clearly into view all the features of rival moral conceptions so that their respective capacities to draw the allegiance of free and intelligent persons under favorable conditions may be fully tested. This allegiance, freely given, and manifested by way of an enduring and effective regulative self-understanding, is what justifies a moral conception.

It is not the allegiance of moral philosophers that is crucial, though it counts. Rather, it is the allegiance of free and intelligent persons capable of competent moral judgment. And this allegiance is claimed only as the final test of whether a moral conception is justified. There is no further metaphysical claim regarding the ground of justification. Moral philosophy inquires into the justification of competing moral conceptions by inquiring into their capacities to draw the relevant allegiance under the appropriate conditions and so to underwrite among persons a shared understanding and experience of their moral relations as mutually intelligible and defensible within the space of reasons. This is its second main task. Explication, of course, is the first.

Moral philosophy emerges, then, as the exercise of practical reason aimed at producing a moral conception capable over time and in light of the empirical facts of human psychology of winning the enduring allegiance of free and intelligent persons and so underwriting for them a certain kind of experience and self-understanding of, and purposeful attention to, their relations with one another. While both ethics as science (explication) and moral theory (systematic comparison of rival moral conceptions) draw on theoretical reason,[7] both are ultimately expressions of and in service of practical reason, which "seeks reasonable means to reasonable ends."[8] Here the reasonable end is a particular kind of experience and self-understanding of, and practical success with, relations among persons. While this end is a reasonable end, there is no further defense of it offered. It is simply available to practical reason as a reasonable end to be pursued by reasonable means. Moral philosophy is the exercise of practical reason in search of the reasonable means.

The foregoing helps to explain Rawls's often noticed reluctance to characterize moral principles as true or false. Truth is not the primary currency of practical reason. Practical reason aims at reasonable means to reasonable ends. Sound moral principles are reasonable means to a reasonable end. Insofar as they do justificatory work, playing a role in the justification of determinate moral judgments, they do so, in the end, not because they are "true," but because they are "reasonable" or because they belong to a "reasonable" moral conception. Of course, it does not follow that there is no sensible point of view from which moral principles or even the "reasoning machine" from which they are generated might properly be regarded as "true" or "false." Nor does it follow that there is no sensible point of view from which determinate moral judgments (justified by reference to reasonable moral principles) might properly be regarded as true or false. Truth may be the language spoken by the voice of (morally competent) conscience. Rawls takes no position on these matters one way or another. His point is simply that as reasonable means to the reasonable end of rendering mutually intelligible and acceptable the relations given or established by our noncontroversially competent moral judgments, moral principles are, on his view, reasonable rather than true. Whether they or the determinate moral judgments they justify are true also, and what it would mean for them to be so, is a question he finds himself disinclined to pursue, apparently failing to see that anything of import hangs in the balance.

With respect to explication, Rawls devotes considerable attention in his dissertation to determining the data to be explicated, noncontroversially competent moral judgments, by ethics as science. He is especially keen to avoid begging any questions. Judgments count as noncontroversially competent if and only if they are the spontaneous, and stable upon reflection, judgments of free and intelligent men or women regarding familiar or standard sorts of cases and made under favorable conditions (not the result of obvious bias, or manipulation, or propaganda; with access to and understanding of relevant information, etc.). So-called moral intuitions generated in response to wild thought experiments far removed from the sorts of cases and circumstances to which human beings have been regularly exposed over successive generations are not included in the data to be explicated. The data to be explicated is, essentially, the public observable manifestation of our moral sentiments as stable, enduring aspects of our nature developed and expressed under favorable conditions.

That there is data of this sort to be explicated by ethics as science is something that we discover. There is no a priori reason to suppose that free and intelligent persons in favorable conditions will converge in their moral judgments to any particular degree. If there is no convergence or no significant convergence, then ethics as science cannot get off the ground and moral philosophy would be stymied before it could get started. But, Rawls insists, there is significant convergence.

But it ought not be overestimated. Rawls explicitly excludes many moral judgments from the data to be explicated by ethics as science. No matter how widespread or confidently held, moral judgments arrived at within contexts characterized by widespread social manipulation, exposure to propaganda, lack of access to or refusals to consider relevant information, relations of subordination or excessive dependency, or similar circumstances, do not constitute good evidence of the shape of the moral sentiments of free and intelligent persons living under favorable conditions. And so Rawls excludes them from the data to be explicated, even if they are widely believed by those holding or making them to be noncontroversially competent. To be sure, he allows that in the end these judgments might be validated as competent moral judgments by the principles given in a successful explication arrived at through ethics as science. But ethics as science cannot start with these judgments, at least not if it aims to show that our moral nature is part of our rational nature freely expressed.

Rawls concedes that this methodological constraint may substantially diminish the data set to work from within ethics as science. The moral theorist would seem to be left with only the noncontroversially competent moral judgments of free and intelligent persons within more or less developed and reasonably just liberal democracies. But given what we know of human history, Rawls notes, we should hardly be surprised to find that we have a limited supply of evidence regarding the exercise under favorable conditions of moral capacities by, or the shape of the sentiments of, free and intelligent persons as such.

This is an important point. Moral philosophy cannot proceed, on his own account, without adequate data. To be adequate, this data must arise against certain background social conditions. These conditions include something like a reasonably just and stable constitutional liberal democracy within which free and intelligent men and women are able to give full and direct expression to their moral sentiments free of distortion by various sorts of familiar material, psychological and social forces. But then moral philosophy depends on our ability to establish, scientifically, as it were, by reference to observable evidence, the existence of these background conditions. Establishing the existence of these conditions scientifically is, like establishing the fact that our moral nature is part of our rational

nature, not something that unfolds within the space of facts, events and causes. It unfolds within the space of institutions, actions and reasons. But it is scientific nonetheless, at least insofar as it answers only to publicly observable evidence. In any event, unless and until we can assure one another, publicly and by reference to publicly observable evidence, that we (or some population) in fact live under the sorts of background conditions that must be present if moral philosophy is to have the sort of data necessary to pursue ethics as science, moral philosophy is stalled in the water. There is, then, a methodological priority to political philosophy over moral philosophy. In 1950, Rawls seems to assume that liberal democracy instantiates favorable background conditions and that contemporary liberal democracies largely are what they appear to be and that we can know this scientifically, as it were, by reference to observable evidence and shared public criteria. One way to read his later work, explored below, is as an attempt to make good on these assumptions.

After devoting roughly the first half of his dissertation to the task of explication and "ethics as science" with respect to judgments regarding the moral worth of character, Rawls turns in the second half to moral philosophy's second basic concern, the issue of justification. He argues that justification ultimately terminates in what he calls "intuitive justification." By "intuitive" he does not mean to refer to the exercise of some special intellectual faculty capable of noninferentially ascertaining true moral principles. Rather he means to refer only to the point at which there is no further point to inquiring after a justification. In moral philosophy, this point is reached when free and intelligent men and women in favorable conditions find themselves not only converging in their noncontroversially competent moral judgments in this or that domain, but also drawn to one and the same explication of that domain and to internalizing it as part of their regulative self-understandings and public self-representations. An explication or moral conception so embraced and internalized is sufficient to secure and maintain mutually intelligible and acceptable, that is, justified, relations among persons, at least within its domain. Once it is in hand, there is no further point to inquiries into moral justification. There is no test of whether the relations among persons are justified beyond whether by reference to the principles of a shared moral conception they prove to be mutually intelligible and acceptable to them as free and intelligent persons under favorable conditions.

Rawls recognizes in his dissertation that it is possible for free and intelligent persons, even under favorable conditions, to err in thinking that a particular moral judgment is noncontroversially competent. We are, individually and together, morally fallible. It is tempting in light of this to think that moral philosophy must fail, for it leaves us powerless to identify such errors. Indeed, it may convince us that we are justified in our erroneous moral beliefs. It is tempting to think that the only way to avoid this possibility is for moral philosophy to inquire into what "really" makes a moral judgment "true" or what "really" makes an action "right," to get beyond justification as Rawls presents it. It is this worry, Rawls suggests, that lies behind the conception of moral philosophy as aimed at the discovery and representation of a moral order fixed prior to and independent of any exercise of practical reason or moral judgment. This is a conception of moral philosophy we should resist, on Rawls's view, not because we have good reasons to believe that there is no such order, but rather because we cannot so easily free ourselves of the burden of our practical reason.

Rawls analogizes this situation to the Supreme Court. Free and intelligent persons under favorable conditions stand to moral judgment in the same relation as the Supreme Court stands to legal judgment. The test of whether a legal judgment is valid or competent is

whether the Court affirms it in light of the principles given by the best explication of other noncontroversially competent legal judgments. Of course, this doesn't mean that the Court's so affirming a legal judgment is what makes it valid or competent. Nor does it mean that the Court cannot err, either in its affirming a legal judgment on review or in the identification of the data – the other noncontroversially competent legal judgments the explication of which yields the principles it draws on when affirming a legal judgment on review. It is tempting, then, as in the case of moral judgment, to suppose that judges or legal philosophers ought to inquire into what "really" makes a legal judgment "true," to somehow get beyond or behind the Court's judgment as the test or criterion of legal validity or competence. But there would be no more to be gained in making this shift in the legal case than in the moral case. Irrespective of one's metaethical or metalegal commitments, there can be no higher or further *test* of whether a legal judgment is valid or competent beyond whether the Court exercising its own best judgment affirms it as such. And there can be no higher or further *test* of whether the Court's best judgment is good enough beyond the Court's own best judgment of the matter. And so on. This does not mean, of course, that the Court is the source of, or infallible in its exercise of, its own authority. It means only that, like practical reason itself, it is the final arbiter of its own authority. There is no further test, only repeated applications of the same final test. Referring to practical reason, Rawls suggests that here lies the deep truth of the saying that you "cannot derive an ought from an is." Practical reason is the final arbiter of its own authority.

Rawls does not restrict the point here about reason as the final arbiter of its own authority to the contexts of legal and moral judgments. He generalizes it to all judgments arising out of the exercise of reason, whether practical or theoretical. There is no higher or further standard of justification within science, for example, beyond what free and intelligent persons engaged in inductive theory confirmation (in the sciences) affirm and internalize (as a reasonable means to their reasonable ends) as an explication of the judgments in these areas they take to be noncontroversially competent. And there is no higher or further standard within logic beyond what free and intelligent persons engaged in deductive inference (in philosophy, mathematics, logic, etc.) affirm and internalize (as a reasonable means to their reasonable ends) as an explication of the judgments in these areas they take to be noncontroversially competent. In each case, reason answers only to itself. And importantly, even when the judgments under examination arise out of the exercise of theoretical reason, as in the cases of inductive and deductive logic, it is ultimately practical reason that has the final word. For the criteria of competent judgments in these cases will be given by an explication of noncontroversially competent judgments in the domain that free and intelligent persons working in that domain accept and internalize as a reasonable means to their reasonable end of mutual intelligibility and justification therein. Remarkably, Rawls held these views while still committed to a roughly Millian utilitarianism. The evidence of Kantian or Wittgensteinian influences is thin. Kant's *Metaphysics of Morals* and *Critique of Practical Reason* are included in the dissertation's bibliography but are not examined in detail in the text. Wittgenstein does not appear in the bibliography.

Here lies, I suggest, the root of Rawls's idea of reflective equilibrium. An explication of competent moral judgment that is effectively regulative within a person (because it draws her allegiance and so belongs to her internalized self-understanding) brings her particular judgments and more abstract theoretical beliefs in line. An explication widely shared by persons in this way brings them, at least with respect to the relevant particular judgments

and more abstract theoretical beliefs, in line with one another. Persons so aligned, and knowing publicly that they are so aligned with one another this way achieve mutually intelligible and justifiable relations with one another as rational social beings. That is, they are justified, or aligned, to one another. Of course, persons make competent moral judgments across many domains – in particular, the domains over which the four exercises of reason, two theoretical and two practical, range. Persons possessed of a shared and effectively regulative explication of competent judgment across all these domains are fully justified, or aligned, to one another as rational social beings. They share as persons a fully intelligible and mutually acceptable or at least reasonable social world. Rawls would eventually characterize these relations in terms of wide, full, general reflective equilibrium. When this state is reached, further inquiry into justification lacks point or purpose. There is no further test available to us. And ultimately it is practical reason that gives and grades the test.

It is here, all the way back in Rawls's dissertation, that we find the roots also of his later constructivism. If complete justification consists in full, wide, general reflective equilibrium, then a, perhaps the most, reasonable way to proceed in working up a reasoning machine capable of explicating noncontroversially competent judgments across one or more domains will be to organize it so that its parts reflect key elements of a familiar normative self-understanding likely to draw the allegiance of free and intelligent persons under appropriate background conditions. For example, a reasoning machine that models a familiar normative self-understanding of democratic citizenship, taken as the basic political office or relationship among persons within a society, and that in turn generates principles capable of explicating (and extending the range of) noncontroversially competent political judgments, would presumably stand a good chance of drawing the allegiance of persons over time and so contributing to their achieving the sorts of relations contemplated by wide, full, general reflective equilibrium.

5. From Moral Philosophy to Democratic Theory

Rawls evidenced some interest in politics and political philosophy as early as his undergraduate thesis. There he is critical of the social contract tradition, taken in its contractarian rather than contractualist form, and insistent that the problem of politics is not the reconciliation of the individual and the social, for personality and community are mutually interdependent, but rather the identification and institutional management of the various forms of "sin" that threaten both personality and community. Ultimately, of course, in the undergraduate thesis Rawls takes the view that human beings are fundamentally corrupt and incapable of achieving and maintaining personality and community without God's grace. And politics and political philosophy presumably must accept this limitation. Rawls's later work aims at a more hopeful, a reasonably hopeful, view of what human beings are capable of, both morally and politically, on their own.

Rawls's interest in politics and political philosophy is evidenced again by his discussion of the "cash value" of "ethics as science" in his 1946 paper, written early in his graduate studies and discussed above. It is further evidenced in later papers written in graduate school and in his 1950 dissertation. Rawls begins his dissertation emphasizing the links between his inquiries in moral philosophy to the possibility and justification of democracy. What had in 1946 been simply a closing remark about the "cash value" of ethics as science had become by 1950

a kind of organizing and motivating principle for his work. Rawls notes that citizens and officials very often disagree in their basic political judgments. These disagreements appear often to be, at their root, moral disagreements, disagreements about the demands of justice or right on their interactions and on their institutions. In a democracy, citizens and officials resolve these disagreements by voting, by exercising the authority of their political office. Or at least they do so when there is a felt need for collective action and the disagreements stand in the way of their acting collectively. So much is obvious.

But under what conditions do they each and all have good reason to acquiesce in collective action determined by voting? Under what conditions, for example, would citizens in the late 1940s or early 1950s each and all have good reason to acquiesce in the outcome of a democratic process concerned to address the civil rights of African-Americans. This question was undoubtedly on Rawls's mind. Most generally, the matter must be one of reasonable rather than merely simple disagreement, to use the terminology I introduced earlier. Citizens generally have no standing reason, certainly no moral reason, to acquiesce to the outcome of a vote taken to resolve a simple disagreement, a disagreement to which one or both parties bring an unreasonable view, a view unintelligible (irrational or unreasonable) in light of the principles that explicate noncontroversially competent judgments in the relevant domain. If citizens generally have a moral reason to acquiesce to the outcome of democratic processes, it is because those processes are deployed to resolve their reasonable moral (and other) disagreements over matters requiring their collective action. But then they must have some shared criteria for distinguishing reasonable from simple moral disagreements. Here moral philosophy can contribute to democratic theory. Evidencing his interest in politics and political philosophy, Rawls intends his dissertation to make, among others, a contribution of this sort. Setting out the moral principles that underwrite competent moral judgment and so demarcate the boundary of reasonable disagreement in a particular domain is a contribution to not only moral but also political philosophy or democratic theory.

This interest in contributing to political philosophy or democratic theory is signaled at the start of his dissertation. Rawls notes that (mid-century) liberal democracies are plagued by two moral outlooks, widespread but fundamentally at odds with liberal democracy. The first he refers to as positivism. Positivist views reduce morality to facts or settled patterns, natural or conventional, without any reference to the authority of practical reason over issues of justification. The positivist might do something that looks like Rawls's explication or ethics as science in order to move beyond mere description to the prediction and explanation of moral phenomena. But the positivist either pursues explication or ethics as science within the space of causes rather than reasons, or treats justification as no more than proof via an exercise of theoretical reason, ignoring the final authority of practical reason to adopt any particular candidate explication as a reasonable means to its reasonable ends. In either case, the positivist does not seek from "ethics as science" an explication of competent moral judgment within the space of reasons the justification of which is determined by whether free and intelligent persons are drawn to adopt it as a reasonable means to their reasonable end of relations of mutual intelligibility and justification within moral community. For the positivist, moral principles constitute facts of one sort or another discoverable and justifiable by theoretical reason. For Rawls, Hume qualified as a positivist. So too did the Scandinavian legal realist Axel Hagerstrom, whom Rawls had read carefully, and the German legal positivist Gustav Radbruch. So too further did the British emotivist A.J. Ayer, and his American analogue, Charles Stevenson.

The second sort of view Rawls identifies as inconsistent with a viable democratic polity is authoritarian. Authoritarian views "assert that ethical principles must be taken on authority, or posited by an act of faith, or at least presupposed." Divine command theories are paradigmatically authoritarian. But so too are those that posit moral principles justified by a transcendental metaphysical necessity discoverable through the theoretical exercise of reason alone.

Positivist and authoritarian views subordinate practical to theoretical reason. And they provide no fertile ground from which to draw a compelling vision of democratic politics capable of winning the enduring allegiance of free and intelligent persons. Rawls characterizes both positivist and authoritarian views as "appeals to exalted entities." The exalted entities he has in mind include God, nature, history, established conventions, all manner of self-proclaimed elites and authorities, theoretical reason, and metaphysics.[9] Moral views built up around appeals to exalted entities subordinate the authority of practical reason to something taken as fixed independent of and prior to any exercise of practical reason. This is a view that Rawls rejects. He insists that practical reason bows before no authority save its own, not even the authority of theoretically demonstrable transcendental or metaphysical necessity. There are no moral principles beyond rational criticism and correction in the face of practical experience and no standard of justification higher or further than the allegiance of free and intelligent men and women. On this, he is quite explicit.

But apart from their rejection of what he affirms, namely the priority and authority of practical reason, why should Rawls worry about positivist and authoritarian views of morality? How do they pose a threat to democracy? The tenor of Rawls's work as a graduate student, including his dissertation, suggests that the answer to these questions is that Rawls suspected that those who affirm positivist or authoritarian views of morality were likely to prove too easily tempted to a kind of dogmatic fanaticism within or apathetic withdrawal from democratic politics. Democratic politics is treated as either a means to their end of vindicating an exalted entity or little more than a modus vivendi or necessary evil. In either case it is hardly the sort of thing to which one pledges full and final allegiance, the sort of allegiance that must be pledged if democracy is to survive in the long run in a world likely to continue to generate threats to it, new Naziisms, new Stalinisms, etc. And so Rawls ventures in his dissertation to contribute to democratic theory, or political philosophy more generally, by taking a step toward an alternative, nonpositivist, nonauthoritarian, yet still fully cognitive, view of morality, one more congenial to a genuinely viable democratic politics. But the contribution here is still one made from within moral philosophy.

It is not until his year at Oxford, 1952–1953, as a Fulbright Fellow that Rawls begins to contemplate a more direct contribution to democratic theory or political philosophy more generally, one made from within, as it were. He departed for Oxford still working toward extending the framework of his dissertation, which focused on judgments regarding the moral worth of character, to include also judgments of value and right action, all within the broadly Millian utilitarian framework with which he was sympathetic at the time. He returned to the United States at work on what would emerge as a theory of justice. I want to conclude this essay with some comments on this transition.

During his year at Oxford Rawls spent a good deal of time thinking about Mill's various precepts of justice. These function as moral principles available to citizens to establish the reasonableness or competence of, to render intelligible and justify to one another, their political views on a wide range of issues. Of course, the precepts will permit reasonable disagreement.

Even if only one precept applies to an issue, citizens may still reasonably disagree over the particular judgment it most supports. But often more than one precept will apply to an issue, so that precepts must be weighed and balanced (or lexically ordered) for the case at hand in order to arrive at a particular judgment. And here too citizens may reasonably disagree.

Now it presumably lies within the political authority of citizens, the authority of their office in a democracy, to resolve these reasonable disagreements by voting. If citizens restrict their voting to issues requiring collective action with respect to which they reasonably disagree, and they vote only for reasonable positions, then a compelling case may be made (and will be much more easily made on the sort of nonpositivist, nonauthoritarian, cognitivist view of morality for which Rawls argues in his dissertation) for the claim that they have a good moral reason to acquiesce to the outcomes of democratic processes. But if they do not so restrict their voting, making such a case seems difficult at best. The question Rawls finds himself contemplating, then, is whether Millian utilitarianism, with its highly plausible precepts of justice, is capable of serving as a public criterion of reasonable political disagreement eligible for democratic resolution. For if it is not, then the particular moral theory he is at work on will prove, in the end, inadequate to the needs of democratic theory and fail as a contribution to political philosophy.

Rawls worries that Millian utilitarianism is not up to the task. And the reason he worries is straightforward. Consider a case in which more than one precept of justice applies so that precepts must be weighed and balanced (or lexically ordered) in light of the particularities of the case in order to arrive as a reasoned determinate judgment. There is no criterion, other than the principle of utility, to which to appeal to establish the reasonableness of any particular weighing and balancing (or lexical ordering) of the precepts. But the principle of utility places no principled constraint on the amount of information relevant to its application. Of course, every application will necessarily draw on information less than the total amount of available and potentially relevant information. Without a further criterion for distinguishing reasonable from unreasonable applications of the principle, or selections of informational input, it looks like the range of reasonable applications of the principle of utility to weigh and balance (or lexically order) Mill's precepts of justice is virtually unlimited. For any claim about how to weigh and balance (or lexically order) the relevant precepts for a particular case there will be an application of the principle of utility capable, with the right informational inputs, of vindicating its reasonableness. But this leaves the democratic authority of citizens unlimited. And surely there is no compelling moral reason to acquiesce to the outcomes of democratic processes in which citizens exercise unlimited authority or in any case cannot publicly establish for one another that they are doing otherwise. If citizens are to acquiesce willingly to the outcomes of democratic processes, those processes must involve and be publicly vindicable as the exercise of limited democratic authority. Of course, the limits may have no deeper ground than the normative self-understandings at work in wide, full and general reflective equilibrium among citizens.

Here Rawls seems quickly to recognize three key points, and they together put him on the path to what would be a theory of justice. The first is that while democratic authority must surely be limited to the resolution of reasonable disagreements, it does not extend to the resolution of all reasonable disagreements. Citizens reasonably disagree over which religion to affirm and embrace. Yet they have no standing general reason to acquiesce in collective action determined by a vote to resolve the disagreement over which religion should be affirmed and embraced. Indeed, they have a standing general reason not to do so. This is something like

a provisionally fixed point within democratic self-understandings. Only certain reasonable disagreements are eligible for democratic resolution.

But which ones? Presumably only those that stand in the way of necessary or desirable collective action for the common good. But how are these notions to be understood? When is collective action necessary or desirable? And how do we give content to the idea of the common good? Even if these questions permit of more than one reasonable answer, we need criteria for distinguishing reasonable from unreasonable answers. This brings us to the second point. To contribute to democratic theory or political philosophy generally a moral theory must answer these questions. The answers given, whether in the form of the idea of the basic social structure as the logically first collective action in which citizens must necessarily engage, or in the form of the priority of certain basic liberties as a necessary feature of any reasonable conception of the common good, need not have any deeper ground than the normative self-understandings at work in wide, general and full reflective equilibrium among democratic citizens.

Taking the first two points together, we can say that Rawls saw the need to provide criteria by which to pick out reasonable disagreements over necessary or desirable action aimed at the common good. These criteria specify the limits of the authority citizens wield in a democracy. They may ultimately be constructed out of widely shared normative self-understandings. And they need pass no justificatory test other than winning the allegiance of free and intelligent citizens standing in relations of wide, full and general reflective equilibrium. But this brings us to the third point. For they must also prove sufficient, in light of observable evidence, to validating objective judgments as to whether any particular citizen is in fact acting within the scope of her authority when she advances or votes for a particular position on a political issue. Citizens and officials must be able to assure one another by reference to these criteria, in light of publicly observable evidence, that when they take official political action they are in fact acting within the scope of their authority and not illegitimately using it to advance ends they are not authorized to advance (for example, something other than the common good or their private good as part of the common good). In the absence of such assurances there remains always the possibility that what appears to be democracy is in fact counterfeit. This possibility is corrosive of the public trust upon which democracy depends. Without this public trust, democracy is at risk of collapsing into a kind of generalized prisoner's dilemma. Unwilling to be played by others as a sucker or chump, and incapable of assuring themselves that they are not being so played, citizens are willing to do no more than pretend at the authority of their office, pretend to be advancing reasonable views about necessary or desirable collective action aimed at the common good, all the while seeking to advance only ends much narrower.

During his year in Oxford, Rawls appears to have become convinced that democratic theory or political philosophy must take up these questions and establish shared public criteria for the normative structure or the authority of citizenship in a democracy. And he found himself uncertain as to whether the Millian utilitarianism he otherwise favored (over available versions of classical utilitarianism) could be developed so that it was fully up to the task. Intuitionist moral views seemed an unpromising alternative, since they offered no criterion for determining a reasonable weighing and balancing or lexical ordering of precepts of justice for any particular case other than that the person judging the case reports that she sincerely judged the matter thus. As with the principle of utility, the intuitionist's criterion for determining a particular application of the precepts of justice to be reasonable leaves the field virtually unlimited. To establish the reasonableness of any particular application of the precepts of justice I

need only say that I sincerely judge it to be reasonable. There is, then, no limit to the scope of my democratic authority other than the limit I choose to place on it. Which is to say, there is no limit.

It is instructive and illuminating, I think, to view what emerged as *A Theory of Justice* as originating in the search for a viable shared public understanding of the normative structure of citizenship in a democracy, an understanding capable of sustaining objective judgments as to whether a political disagreement was in fact a reasonable disagreement properly settled by democratic procedures or was merely a simple or unreasonable disagreement masquerading as a reasonable disagreement. While Millian utilitarianism and intuitionism were clearly moral views more congenial to democracy than positivist or authoritarian views, they seemed to Rawls by the early 1950s less than ideally equipped to deliver the sort of shared public understanding of citizenship necessary to the long-term stability and vitality of a democracy. *A Theory of Justice* takes a fresh start, taking a constructivist and contractualist approach to the problem. Rawls's two principles of justice, already lexically ordered and more or less easily applied to observable evidence, establish the nature and scope of the authority citizens wield, and so mark the field of reasonable political disagreement, whether at a constitutional convention, a legislative session, a judicial decision, or a vote at the ballot box. The first principle affirms each citizen's equal claim to a system of basic liberties fully adequate to the development and exercise of her moral capacities. The second affirms each citizen's equal claim to a basic social structure within which the different offices or positions, with their different powers and responsibilities and different economic rewards, arising out of a structurally maintained division of labor are, first, open to all on terms of fair equality of opportunity and, second, such that there is no alterative scheme of offices and positions, no alternative division of labor, with a superior least desirable social position. (The social position occupied by the typical unskilled laborer is the least desirable social position, in terms of lifetime expected income, arising out of the division of labor structurally maintained in most advanced industrial democracies. This is just only if there is no alternative division of labor with a least desirable position superior to that of today's typical unskilled worker.) In a democracy an exercise of official political power is counterfeit if it is inconsistent with either of these two principles or their lexical ordering. The two principles, conjoined with the idea of the basic structure as the first subject of justice, express the criteria by which citizens can validate a particular issue as one within the scope of their political authority qua citizen. Rawls's hope, of course, is that the two principles and the original position (reasoning machine) used to generate them not only explicate what we take to be noncontroversially competent judgments regarding the nature, scope and reasonable exercise of the authority vested in democratic citizenship, but that they do so in a way, or as part of a larger political moral conception, capable of drawing our enduring allegiance as free and intelligent men and women. If he is right, then justice as fairness is a, perhaps the most, reasonable means to our reasonable end of realizing and maintaining as free equals mutually intelligible and justifiable political relations. For this end, of course, there is no argument. There is just the power of a picture – a picture, reasonable to be sure, of what we are and what we might be.

Notes

Research for this essay was supported by, and I express my gratitude for, a Fellowship from the National Endowment for the Humanities.

1 Rawls famously often analogized his life's work to a single painting – his attempt to set out the single vision that had captured his attention for a half-century, even as he struggled to bring it into focus, attended to different elements at different times, and periodically found himself adjusting this or that feature in an attempt to get it "just right."

2 I discuss the undergraduate thesis at some length in Reidy 2010.

3 This is a point that suggests the influence of Wittgenstein. But, interestingly, throughout the later 1940s and through his dissertation in 1950, Rawls draws on figures associated with "scientific philosophy" and the Vienna Circle other than Wittgenstein.

4 By the early 1950s, Rawls began to worry about this feature of utilitarianism, at least if utilitarianism is to serve the social role of morality by supporting a shared public criterial understanding of the competent exercise of our moral faculties. Because there are no principled limits on the amount of relevant information plausibly fed into the utilitarian "reasoning machine" and because the machine's output, the theorems or moral principles it generates, will inevitably vary with the information given as input, which must always be less than all possibly relevant information, the theory's ability to serve as a reasonable means to our reasonable end of rendering our competent moral judgments mutually intelligible and justifiable seems to depend on some accidental, unexplained and probably unlikely agreement as to the portion of all relevant information to be fed into the reasoning machine. By the late 1950s or early 1960s, Rawls had concluded that this problem was sufficient to render even the most plausible sort of utilitarianism, some version of Millian utilitarianism of the sort he'd been developing, unworkable as the theoretical framework for a public theory of justice within a democracy.

5 In commenting on his proposed "imperative utilitarianism," Rawls notes that there is no reason to think that a sound scientific understanding of the effective exercise of our moral capacities or of our competent moral judgments will contribute to our more reliably acting as required by morality. There remains, he notes, the problem of the "radical evil" of human nature, a problem that cannot be overcome save through divine grace. This is a remarkable note for several reasons. First, it indicates that in 1946 Rawls was not yet fully out from under the shadow of the religious views expressed in his undergraduate thesis. Notwithstanding his recollection late in life of having abandoned his faith during or very shortly after the war, the evidence suggests that Rawls's prewar theistic commitments were slowly altered and abandoned over a period of some 10 years following the end of the war. In 1954, Rawls taught a course in Christian Ethics at Cornell. He was then still thinking very seriously about Christianity and was still in the process of finding his way to the "nontheistic" orientation on which he would eventually settle by the late 1950s or early 1960s. Second, it indicates the extent to which Rawls conceived of ethics, at least in 1946, as aimed primarily at our self-understanding as persons and not at making or improving us as persons. Over time, he would come to think of moral philosophy as making a practical contribution to our complete realization, including our improvement or education, as persons.

6 This, of course, would change in the early 1950s. The change would be announced in the paper "Two Concepts of Rules," which Rawls regarded as clearing the way for the addition of institutions and practices to the list of subjects properly evaluated by moral judgment and so properly covered by a complete moral theory.

7 Rawls emphasizes that these two inquiries at the heart of moral philosophy – one into explication and the other into justification and motivation – may be intelligently carried out without much attention being paid to the sorts of metaphysical, metaethical, and linguistic inquiries that seemed so often to distract, even to dominate the attention of, moral philosophers mid-century. One might add today that they may be so carried out without much attention to many of the empirical inquiries that seem so often to distract, even dominate, the attention of moral philosophers today.

8 Rawls characterizes reasonable ends as activities which are comprehensively satisfactory for the person whose activities they are and inclusively harmonious with the comparable activities of other persons within the community to which the person belongs.

9 Rawls includes here the sort of nonnaturalist "realism" urged by G.E. Moore. Rawls agrees with Moore in holding that "good" cannot be analyzed – in the sense of "synonymy meaning," an analysis that would preserve truth values through substitution – in purely naturalistic terms. It is, in that sense, a nonnatural or moral property. But Rawls rejects the claim that this linguistic fact entails any further fact, metaphysical or otherwise, discoverable through the exercise of theoretical reason to which practical reason must acquiesce.

Works by Rawls, with Abbreviations

A Brief Inquiry into the Meaning of Sin and Faith, with "On My Religion" (*BI*), ed. Thomas Nagel. Cambridge, MA: Harvard University Press, 2009.

"A Brief Inquiry into the Nature and Function of Ethical Theory" (1946), Rawls Archive: Box 3, Accession 14990, Pusey Library, Harvard University.

Collected Papers (*CP*), ed. Samuel Freeman. Cambridge, MA: Harvard University Press, 1999.

"Ethical Rationalism" (late 1940s). Rawls Archive, Box 2, Accession 14990, Pusey Library, Harvard University.

"A Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments of the Moral Worth of Character." PhD dissertation, Princeton University, 1950.

A Theory of Justice (*TJ*), rev. edn. Cambridge, MA: Harvard University Press, 1999.

"Two Concepts of Rules," *Philosophical Review* 64(1) (1955): 3–32. Also in *Collected Papers* (20–46).

Other References

Downs, Anthony (1957) *An Economic Theory of Democracy*. New York: Harper & Row.

Kelsen, Hans (1978 [1934]) *The Pure Theory of Law*. Berkeley: University of California Press.

Reidy, David A. (2010) "Rawls's Religion and Justice as Fairness." *History of Political Thought* 31: 309–343.

Schumpeter, Joseph (1942) *Capitalism, Socialism and Democracy*. New York: Harper.