RAWLS' CONCEPT OF REFLECTIVE EQUILIBRIUM AND ITS ORIGINAL FUNCTION IN *A THEORY OF JUSTICE*

JOHN MIKHAIL*

John Rawls' concept of reflective equilibrium[1] has been the subject of considerable interest and debate. Its use in moral philosophy has drawn criticism from many philosophers and legal scholars[2] and has been resourcefully defended by others.[3] Rawls' concept originally derived from an influential account of the philosophical method of justifying principles of inductive inference.[4] As such, many commentators have assumed that the concept implies a close nexus between the methods of ethics and empirical science, and thus may shed light on the nature of moral truth, justification, and objectivity.[5] Reflective equilibrium also plays a significant role in the analogy Rawls draws in *A Theory of*

[*] Professor of Law, Georgetown University Law Center. This Essay is a slightly modified and expanded version of the discussion of reflective equilibrium that appears in my doctoral dissertation on Rawls' linguistic analogy. *See infra* note 19. I wish to thank the editors of the *Washington University Jurisprudence Review* for the valuable opportunity to publish this version in their journal and for their excellent editorial work in bringing the Essay to publication. Thanks also to the many scholars who, over the course of nearly two decades, have helped to improve my understanding of reflective equilibrium and its original function in *A Theory of Justice*, including most prominently David Brink, Noam Chomsky, Tom Grey, Gilbert Harman, Terry Irwin, Christine Korsgaard, David Lyons, Matthias Mahlmann, Richard Miller, John Rawls, Henry Richardson, Tim Scanlon, Phillipe Schlenker, Peter Singer, Larry Solum, Liz Spelke, Jason Stanley, Nicholas Sturgeon, Josh Tenenbaum, Alec Walen, Leif Wenar, and Allen Wood. Finally, I wish to thank Sean Kellem for his outstanding assistance in helping me to prepare this Essay for publication.

[1] *See* JOHN RAWLS, A THEORY OF JUSTICE 18–22, 46–53 (1971).

[2] *See, e.g.*, RICHARD BRANDT, A THEORY OF THE GOOD AND THE RIGHT (1979); RICHARD A. POSNER, THE PROBLEMATICS OF MORAL AND LEGAL THEORY (1999); R. M. Hare, *Rawls' Theory of Justice*, *in* READING RAWLS: CRITICAL STUDIES ON RAWLS' 'A THEORY OF JUSTICE' 81–107 (Norman Daniels ed., 1989) [hereinafter READING RAWLS]; Peter Singer, *Sidgwick and Reflective Equilibrium*, 58 THE MONIST 490 (1974).

[3] *See, e.g.*, RONALD DWORKIN, TAKING RIGHTS SERIOUSLY (1977); Joseph Raz, *The Claims of Reflective Equilibrium*, 25 INQUIRY 307 (1982).

[4] *See* NELSON GOODMAN, FACT, FICTION, AND FORECAST (1955; paperback, 1983).

[5] *See, e.g.*, DAVID BRINK, MORAL REALISM AND THE FOUNDATION OF ETHICS (1989); Richard Boyd, *How to Be a Moral Realist*, *in* ESSAYS ON MORAL REALISM 181–228 (G. Sayre-McCord ed., 1988).

Justice between moral theory and generative linguistics.[6] The linguistic analogy has been a source of intense controversy in its own right and has led to further speculation about whether Rawls considers moral theory to be an empirical discipline—a branch of psychology—and, if so, whether its claims to normative authority can be maintained.[7]

Unfortunately, scholarly discussions of reflective equilibrium often proceed as if there were a single, uniform notion going under this name. This assumption is mistaken. Rawls' original explanation of how to reconcile the descriptive and normative aspects of moral theory in his early article, *Outline of a Decision Procedure for Ethics*[8] (hereinafter *Outline*), differs significantly from the account of the same problem offered in *A Theory of Justice*, differences that go well beyond the absence of the term "reflective equilibrium" in the former article. Moreover, Rawls' two detailed accounts of reflective equilibrium in Sections 4 and 9 of *A Theory of Justice* are themselves dissimilar in certain respects, although Rawls does not call attention to these differences. Further complicating matters is the distinction Rawls draws implicitly in Section 9 of *A Theory of Justice*, and explicitly in *The Independence of Moral Theory*[9] (hereinafter *Independence*), between *narrow* and *wide* reflective equilibrium. In a pair of important articles published over thirty years ago, Norman Daniels supplied a detailed treatment of this distinction as a means of responding on Rawls' behalf to the criticisms of philosophers such as R. M. Hare, Peter Singer, and Richard Brandt.[10] Yet although Daniels' interpretations of this distinction are often taken to be authoritative, there appear to be significant discrepancies between the original distinction Rawls draws between narrow and wide reflective equilibrium in both *A Theory of Justice* and

[6] *See generally* RAWLS, *supra* note 1, at 46–51, 491.

[7] *See, e.g.*, DWORKIN, *supra* note 3; BERNARD WILLIAMS, ETHICS AND THE LIMITS OF PHILOSOPHY 93-119 (1985); Thomas Nagel, *Rawls on Justice*, 82 PHIL. REV. 220 (1973), *reprinted in* READING RAWLS, *supra* note 2, at 1–16; Hare, *supra* note 2; Raz, *supra* note 3; Singer, *supra* note 2.

[8] John Rawls, *Outline of a Decision Procedure for Ethics*, 60 PHIL. REV. 177 (1951).

[9] John Rawls, *The Independence of Moral Theory*, 48 PROCEEDINGS AND ADDRESSES OF THE AMERICAN PHILOSOPHICAL ASSOCIATION 5 (1975).

[10] *See* Norman Daniels, *Wide Reflective Equilibrium and Theory Acceptance in Ethics*, 76 J. PHIL. 256 (1979); Norman Daniels, *On Some Methods of Ethics and Linguistics*, 37 PHIL. STUD. 21 (1980). Both articles are reprinted in Daniels' book, JUSTICE AND JUSTIFICATION: REFLECTIVE EQUILIBRIUM IN THEORY AND PRACTICE (Cambridge Univ. Press, 1996).

Independence and the account Daniels offers of the same distinction.

The guiding idea of this Essay is that carefully distinguishing Rawls' different accounts and uses of reflective equilibrium in his early writings may help to clarify some of the debates that this concept subsequently generated in the philosophy and jurisprudence literatures. Indeed, I suspect that many of the interpretive controversies surrounding *A Theory of Justice*—for example, whether Rawls' theory is properly understood to be a type of moral anthropology;[11] whether Rawls is a subjectivist[12] or intuitionist[13] about morality; whether reflective equilibrium represents a discovery procedure or a construction procedure;[14] whether the ultimate justification of Rawls' two principles of justice is coherentist or contractualist;[15] and whether Rawls is committed to a coherence theory of moral truth[16]—may appear different once Rawls' varied and sometimes inconsistent accounts of reflective equilibrium are teased apart and examined.

A comprehensive treatment of the role played by reflective equilibrium in Rawls' early writings might begin by carefully examining how Rawls conceives of the relationship between the problems of descriptive and normative adequacy in his PhD Dissertation, *A Study of the Grounds of Ethical Knowledge*[17] (hereinafter *Grounds*), focusing on how Rawls conceives the structure of an ethical theory in *Grounds* to be comprised of two main parts: explication and justification. One might then turn to Rawls' remarks in *Outline* about constructing a decision procedure for ethics, trying to assess the exact aim of his enterprise there, and focusing on how he treats explication both as an empirical inquiry and as best compared to the study of inductive logic. Ideally, one would also examine Nelson Goodman's classic account of induction, to which Rawls refers when first explaining the meaning of reflective equilibrium in *A Theory of Justice*.[18] In doing so, one might pay particularly close attention to Goodman's explanation of why philosophical attempts to provide a logical justification of

[11] Hare, *supra* note 2.

[12] Singer, *supra* note 2.

[13] BRANDT, *supra* note 2.

[14] DWORKIN, *supra* note 3.

[15] *See, e.g.*, David Lyons, *The Nature and Soundness of the Contract and Coherence Arguments*, *in* READING RAWLS, *supra* note 2, at 141–167.

[16] BRINK, *supra* note 5.

[17] John Rawls, *A Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral Worth of Character* (1950) (unpublished PhD Dissertation, Princeton University) (on file with author).

[18] *See* RAWLS, *supra* note 1, at 20.

induction are permanently unsuccessful—hence why philosophers should relax the distinction between justifying induction and describing ordinary inductive practice—and consider its relevance for Rawls' conception of moral theory. Finally, one might also take up the multiple references to both reflective equilibrium and justification in *A Theory of Justice* and Rawls' other early writings and compare them with the concepts of reflective equilibrium and justification presupposed by Rawls' critics, in order to discover whether any of the controversies surrounding these concepts may be recast, or may simply dissolve upon closer analysis.

A comprehensive discussion of this type goes beyond the scope of this Essay.[19] Instead, my aim here is more modest: to explain the meaning and original function of reflective equilibrium in *A Theory of Justice*. To accomplish this objective, I first briefly clarify the technical nature of this concept and then summarize the main contractual argument of Rawls' influential book (Part I). Next, I explain the meaning and function of reflective equilibrium in Sections 4 and 9 of *A Theory of Justice*, calling attention to several apparent and previously unnoticed differences between these two distinct accounts (Part II). I then discuss the distinction Rawls draws implicitly in *A Theory of Justice* and explicitly in *Independence* between narrow and wide reflective equilibrium (Part III). Finally, I discuss and criticize Daniels' influential interpretation of the latter distinction, making plain the differences between Daniels' deliberately non-psychological account of wide reflective equilibrium and Rawls' own partly psychological account (Part IV). Throughout the Essay, my primary purpose is careful exegesis and analysis of what Rawls actually says about reflective equilibrium in *A Theory of Justice* and *Independence*. This effort is a necessary first step in clarifying many of the philosophical and jurisprudential debates that have surrounded the meaning and function of this concept, as well as many debates about the aims, scope, and authority of moral philosophy more generally during the past four decades.

[19] For one effort in this direction, see generally JOHN MIKHAIL, ELEMENTS OF MORAL COGNITION: RAWLS' LINGUISTIC ANALOGY AND THE COGNITIVE SCIENCE OF MORAL AND LEGAL JUDGMENT (forthcoming, Cambridge Univ. Press); *see also* John Mikhail, *Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice'* (2000) (unpublished Ph.D. dissertation, Cornell University) (on file with author) [hereinafter Mikhail, *Rawls' Linguistic Analogy*].

I. THE TECHNICAL NATURE OF REFLECTIVE EQUILIBRIUM AND THE MAIN CONTRACTUAL ARGUMENT OF *A THEORY OF JUSTICE*

The first and perhaps most important thing to notice about reflective equilibrium is that it is not a term of ordinary language, but an invented, technical term—a term of art in the traditional sense. Like all invented terms, therefore, it has whatever meaning its author gives it. In *A Theory of Justice*, Rawls defines reflective equilibrium in Sections 4 and 9 and mentions it in passing on only four other occasions.[20] He also briefly clarifies its meaning in *Independence*. Despite these explicit definitions, it is not uncommon in the literature on Rawls to find commentators investing reflective equilibrium with new and unwarranted meanings of their own. Among Rawls' early commentators, R.M. Hare, Ronald Dworkin, Peter Singer, and Norman Daniels appear to fall into this pattern quite frequently, although many other writers do so as well.

Reflective equilibrium makes its first appearance in Section 4 of *A Theory of Justice*, where it is defined as a hypothetical state of affairs that is reached in the course of attempting to justify the original position by resolving expected discrepancies between our considered judgments and the principles yielded by a candidate description of the initial situation. In Rawls' theory, *original position*, *considered judgments*, and *initial situation* are also technical terms. Accordingly, the Section 4 definition of reflective equilibrium cannot be understood without a clear grasp of the meanings of these concepts. For the latter purpose, a brief summary of the main, contractual argument of *A Theory of Justice* is required.

Rawls' main contractual argument in *A Theory of Justice* may be understood, for our purposes at any rate,[21] as an attempt to propose and defend a specific solution to the problem of normative adequacy (that is, the problem of which moral principles are justified) that is a viable alternative to utilitarianism and intuitionism. Rawls characterizes his proposal as a "workable and systematic moral conception"[22] that is implicit in the contract

[20] *See* RAWLS, *supra* note 1, at 20, 48-51, 120, 432, 434, 579.

[21] Rawls' complete argument in *A Theory of Justice* is rich and complicated, and the brief summary offered here is thus selective in its points of emphasis. Many important themes and distinctions are ignored. My aim is merely to present a suitably clear picture of the essentials of the book's theoretical structure, primarily as they pertain to the meaning and function of reflective equilibrium.

[22] RAWLS, *supra* note 1, at viii.

tradition of Locke, Rousseau, and Kant. As compared with its two main rivals, it better "approximates our considered judgments of justice"[23] and "constitutes the most appropriate moral basis for a democratic society."[24]

Like moral conceptions generally, a conception of justice enunciates a set of rules or procedures by which ethical questions are to be answered and ethical disputes resolved. What singles out a conception of justice, according to Rawls, are the kinds of questions and disputes that fall under its jurisdiction. In its widest sense, justice is ascribed not only to laws, institutions, and social systems, including those within and among nation-states, but also to particular actions, attitudes, and dispositions of individual persons. Rawls, however, limits the scope of his inquiry to a conception of *social* justice. Human society, which is "a cooperative venture for mutual advantage," produces by its collaborative effort a net surplus of advantages and benefits.[25] A conception of social justice provides a set of principles "for choosing among the various social arrangements which determine this division of advantages and for underwriting an agreement on the proper distributive shares."[26]

Rawls distinguishes various *conceptions* of justice from the *concept* of justice, the latter of which may be thought of as "specified by the role which . . . these different conceptions[] have in common."[27] From the concept of justice it is possible to derive certain substantial requirements that just institutional arrangements must satisfy, for example, that "no arbitrary distinctions are made between persons in the assigning of basic rights and duties" and that "the rules determine a proper balance between competing claims to the advantages of social life."[28] However, the mere concept of justice leaves important questions unsettled, such as how to interpret the notions of an arbitrary distinction and a proper balance. For these, a particular conception of justice is required.

Rawls narrows his topic in two more fundamental respects. First, he focuses on a special case of the problem of justice, which he calls "the primary subject of justice" or the "basic structure of society."[29] This problem concerns how the major social institutions—the political constitution and the primary economic

[23] *Id.*

[24] *Id.*

[25] *Id.* at 4.

[26] *Id.*

[27] *Id.* at 5.

[28] *Id.*

[29] *Id.* at 7.

and social arrangements—should assign basic rights and duties and determine the division of the advantages of social cooperation. Second, Rawls limits his attention to *ideal* rather than *non-ideal* theory. Thus, he seeks principles that would govern a "well-ordered society," which is designed to advance the good of its members and to be regulated by a public and publicly accepted conception of justice.[30]

Rawls' main argument in *A Theory of Justice* may be characterized, then, as an argument for a particular conception of social justice, the principles of which are to answer moral questions and resolve moral disputes about the basic structure of a well-ordered society. The complete argument is complex and includes numerous idealizations and simplifying devices. One guiding idea, borrowed from social contract theory, is nonetheless straightforward. Rawls' proposed conception of justice is not deduced from a presumed self-evident or a priori proposition, nor is it arrived at by an analysis of moral concepts or of the meanings of ethical terms. Instead, its principles are the object of a rational choice made by persons in a hypothetical contractual arrangement, and the fact that certain principles and not others would be chosen under such an arrangement constitutes at least *one* argument for their being a solution to the problem of normative adequacy.

Rawls' contract argument, as Lyons usefully labels it,[31] consists of two parts: first, a characterization of an initial contractual arrangement, and second, a conception of justice that would be chosen in that situation. As Rawls observes, the two parts are logically independent. It is possible to object to one part and not the other. Thus, one might reject the particular conditions and constraints Rawls imposes on the initial contractual situation, or alternatively, one might accept Rawls' characterization of the initial situation but argue that the principles he derives from it are invalid.[32]

The bulk of *A Theory of Justice* is devoted to working out and defending a solution to the choice problem presented by this

[30] *Id.* at 4–5, 8–9. Like reflective equilibrium, *ideal theory* and *non-ideal theory* are terms of art for Rawls. His fullest explanation of them in *A Theory of Justice* may be found in Sections 39 and 53. Ideal theory assumes strict compliance with the principles of justice and seeks to determine the principles that would characterize a well-ordered society under these and other favorable circumstances. By contrast, non-ideal theory assumes partial compliance and considers which principles of justice to adopt and apply in these and other less favorable circumstances, taking up such subjects as just war theory and the duty to comply with unjust laws. *See generally id.* at 8-9, 245-6, 351.

[31] *See* Lyons, *supra* note 15, at 150.

[32] RAWLS, *supra* note 1, at 15, 54.

contractual arrangement. As is well known, the conception of justice he advocates comes in two forms, one general and one more specific. The two principles of the specific conception—what Rawls calls the *special conception of justice*—are as follows:

> *First Principle*
> Each person is to have an equal right to the most extensive total system of basic liberties compatible with a similar system of liberty for all.[33]
>
> *Second Principle*
> Social and economic inequalities are to be arranged so that they are both (a) to the greatest benefit of the least advantaged and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.[34]

To understand the contract argument for the special conception of justice, it is important to explain some terminology that Rawls employs in connection with the first part of the argument. Rawls draws a fundamental distinction between the idea of an initial contractual arrangement and of particular interpretations of it. The former, which Rawls names *the initial situation*, may be thought of as a common feature of contract views generally. The initial situation, in turn, admits of many possible interpretations, each of which gains expression in the form of a set of conditions and constraints imposed there. These usually include a description of the contracting parties, such as their motives and nature, the extent to which they are rational, their knowledge of themselves, and so on. It also includes a general description of their deliberative circumstances, such as the historical information available to them and the formal constraints under which they deliberate.[35]

The name that Rawls gives to his own preferred interpretation of the initial situation is *the original position*. Among its most salient features are the following. Most famously, the parties in the original position are presumed to operate behind a "veil of ignorance" that temporarily prevents them from knowing their identities, their natural talents and skills, their conception of the good, and their social circumstances generally. The deliberators do, however, possess full knowledge of all the general facts and

[33] *Id.* at 250.

[34] *Id.* at 83.

[35] *Id.* at 18, 121.

scientific laws relevant to the determination of principles of justice they must make.[36] Moreover, they are presumed to be rational insofar as they are able to choose the most effective means to given ends and—despite the absence of information about their particular ends—to rank their alternative ends self-interestedly in accordance with a postulated preference for more, rather than less, primary social goods.[37] They are also presumed to be mutually disinterested.[38]

A number of other important features characterize the original position as well. For instance, one of the objective circumstances of the parties is moderate scarcity: demand outruns supply, and resources are neither abundant enough to make cooperation superfluous, nor scarce enough to make it futile.[39] As for candidate principles presented to the parties, they must meet certain formal requirements even to be eligible for consideration. Rawls arranges these requirements under five headings: generality, universality, publicity, order, and finality. Taken conjointly, these five conditions exclude several variants of egoism as serious candidates for a suitable conception of justice. However, the traditional versions of utilitarianism and intuitionism do satisfy these conditions, and hence these formal constraints do not prejudge the issue between Rawls' theory and its main rivals.[40] Finally, Rawls assumes that it is rational for the parties in the original position to adopt a "maximin" rule, which instructs them to prefer principles whose worst possible outcome is superior to that of alternative principles.[41]

According to Rawls, the original position is "the most philosophically favored interpretation" of the initial situation.[42] By this he means that the restrictions imposed on the initial situation are more reasonable and widely accepted than any alternative set of conditions, not only from the standpoint of rational choice but also from the standpoint of moral theory. In other words, the conditions and constraints embodied in the original position are not only reasonable, but they are also morally defensible in that "the principles that would be chosen, whatever they turn out to be, are acceptable from a moral point of view."[43] The original position

[36] *Id.* at 136–38.

[37] *Id.* at 14, 142–43.

[38] *Id.* at 13, 127.

[39] *Id.* at 127.

[40] *Id.* at 130–36.

[41] *Id.* at 152–53.

[42] *Id.* at 18, 122.

[43] *Id.* at 120.

thus constitutes a model of what Rawls calls *pure procedural justice*: the principles it generates are normatively valid principles of justice because they are the outcome of a procedure that has been characterized in such a way as to give them moral force.[44] In particular, the initial situation is interpreted so as to respect the freedom, equality, and rational self-interest of the contracting parties. This feature of the original position is captured by the name Rawls gives to his conception of justice: *justice as fairness*. It conveys the idea that the principles of justice are the product of an agreement that is itself fair.[45]

It is this aspect of procedural fairness that is thought to lend the original position its justifying force. Rawls contends that the intuitively fair and plausible nature of the conditions and constraints embodied in the original position constitutes a provisional justification of the set of principles chosen by the people occupying that position.[46] The justification is only *provisional* because there is more to establishing the procedural fairness of the original position than showing that the conditions imposed there *seem* fair and reasonable. Additionally, one defends the fairness of a particular interpretation of the initial contractual situation by determining whether

> the principles which would be chosen match our considered convictions of justice or extend them in an acceptable way. We can note whether applying these principles would lead us to make the same judgments about the basic structure of society which we now make intuitively and in which we have the greatest confidence; or whether, in cases where our present judgments are in doubt and given with hesitation, these principles offer a resolution which we can affirm on reflection.[47]

To test a proposed description of the initial situation, then, we measure the consequences of its principles against our pretheoretical moral intuitions about the basic structure of society as it applies to particular cases or controversies. The intuitions in question are what Rawls calls "our considered convictions of justice," that is, those judgments "we now make intuitively and in

[44] *Id.* at 85, 136.
[45] *Id.* at 11–12, 136.
[46] *Id.* at 18.
[47] *Id.* at 19.

which we have the greatest confidence."[48] If, on reflection, the principles chosen in the contractual situation can be shown to cohere with our considered judgments about these same cases, then we possess further evidence of the fairness of our proposed interpretation of the initial situation and further grounds for having satisfied the requirements of pure procedural justice.

II. THE MEANING OF REFLECTIVE EQUILIBRIUM IN *A THEORY OF JUSTICE*

Thus far I have not said anything specific about how reflective equilibrium fits into this picture. It first enters Rawls' discussion in Section 4 in the explicit service of the contract argument, as a state of affairs that is reached in the course of resolving expected discrepancies between our considered judgments and the principles generated by a candidate description of the initial situation, in an effort to justify the original position. Later, in Section 9, Rawls characterizes reflective equilibrium somewhat differently, not merely as a state of affairs achieved in the course of justifying the original position, but also as a state of affairs that results after a person has been given the opportunity to evaluate and reflect on competing theoretical descriptions of her sense of justice and has either revised her initial judgments or held fast to them. Significantly, Rawls does not emphasize the difference between these two accounts of reflective equilibrium. Instead, the Section 9 account is presented as if it were a resumption and elaboration of the earlier discussion in Section 4. Yet, as we shall see, there appear to be at least superficial differences between these two accounts.

A. The Section 4 Account of Reflective Equilibrium

We can begin to understand the Section 4 account of reflective equilibrium by recognizing the role that this concept plays in support of the contract argument. According to the contract argument, the two principles of justice are justified because they would be chosen in a suitably characterized contractual situation.[49] This justificatory strategy, as we have seen, is procedural: in the absence of an independent criterion for justice, we construct a procedure whose fairness ensures an outcome that is likewise just

[48] *Id.* at 19–20.

[49] *Id.* at 11, 118–19, 136, 579.

or fair. The original position is itself justified as the most suitable interpretation of the initial situation in two ways: first, by grounding its conditions in intuitively plausible and commonly shared presumptions, and second, by testing its consequences against our considered judgments of justice.[50]

In Section 4, reflective equilibrium is introduced as a state of affairs toward which we strive in attempting this second sort of justification of a proposed description of the initial situation. Presumably, when we test a proposed description, we discover discrepancies between our considered judgments, on the one hand, and the consequences of the principles generated by the description under consideration, on the other. We advance toward reflective equilibrium by attempting to resolve these differences.

Here it is helpful to distinguish our motivation for advancing toward reflective equilibrium from our manner of doing so. According to Rawls, the reason *why* we want equilibrium, and why we don't simply settle for whichever principles a plausible set of contractual conditions generates, is that we may have stronger convictions about certain specific issues of social justice than we do about what constitutes a fair contractual arrangement. For example, consider our firm convictions, which Rawls cites, that religious intolerance and racial discrimination are unjust.[51] Because of the strength of these convictions, we may reasonably be sceptical of an otherwise plausible description of the initial situation whose consequent principles fail to honor them. If our convictions are strong enough, we may go so far as to invalidate such a description. Therefore, the *method* of advancing toward reflective equilibrium is to modify either the description or the considered judgments, for the latter, like the former, are merely provisional fixed points that may be changed. We go back and forth like this, sometimes conforming the description to our judgments, sometimes our judgments to the description and its consequent principles, until at last we "find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted."[52] When this happens, reflective equilibrium is reached, and the description constitutes a temporarily stable conception of the original position. The conception is only temporarily stable because, as its name implies, reflective equilibrium is a state of affairs that is always subject to further

[50] *Id.* at 86, 111, 136.

[51] *Id.* at 19.

[52] *Id.* at 20.

reflection and change. Nonetheless, according to Rawls, once the state of reflective equilibrium is reached, the claim that the original position is the philosophically most favored interpretation of the initial situation is then fully justified—both because the constraints it imposes are independently plausible and because its principles conform to our considered judgments of justice. Since the original position is fully justified, by transitivity the principles of justice it yields are also justified. Rawls describes the entire process as follows:

> In searching for the most favored description of this [initial] situation we work from both ends. We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield a significant set of principles. If not, we look for further premises equally reasonable. But if so, and if these principles match our considered convictions of justice, then so far well and good. But presumably there will be discrepancies. In this case we have a choice. We can either modify the account of the initial situation or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium. It is an equilibrium because at last our principles and judgments coincide; and it is reflective since we know to what principles our judgments conform and the premises of their derivation. At the moment everything is in order. But this equilibrium is not necessarily stable. It is liable to be upset by further examination of the conditions which should be imposed on the contractual situation and by particular cases which may lead us to revise our judgments. Yet for the time being we have done what we can to render coherent and to justify our

> convictions of social justice. We have reached a conception of the original position.[53]

For our purposes, three points about how Rawls characterizes reflective equilibrium in this passage are especially worth emphasizing. First, reflective equilibrium is not, strictly speaking, a "method" or "technique," as Dworkin and other writers frequently describe it,[54] but rather *a state of affairs*. Specifically, it is the state of affairs that exists once a description of the initial situation has been reached "that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted."[55]

Second, the word "reflective" appears to play the same role in Rawls' definition of reflective equilibrium as the word "generative" plays in Chomsky's notion of a generative grammar. In Chomsky's framework, "generative" simply means "explicit."[56] In other words, a linguistic grammar qualifies as a *generative* grammar when its principles can be explicitly stated by the linguist in a format suitable for serving as the premises of a derivation. This is essentially the same meaning Rawls assigns to a *reflective* equilibrium. It refers to the fact that once we have found a correct description of the initial situation, "we know to what principles our judgments conform and the premises of their derivation."[57]

Third, Rawls holds that any proposed description of the initial situation is merely provisional and open to modification as a result of further investigation, hence not necessarily stable. This emphasis on the provisional nature of the original position is important: it implies that it is always an open question whether the initial situation has been accurately characterized and, therefore, whether our convictions of social justice are justified. Rawls emphasizes the provisional or contingent character of reflective equilibrium (understood as the state of affairs that obtains when the initial situation has been accurately characterized) throughout

[53] *Id.* at 20-21. Rawls adds the following footnote to this paragraph, the significance of which I return to below: "The process of mutual adjustment of principles and considered judgments is not peculiar to moral philosophy. See Nelson Goodman, *Fact, Fiction, and Forecast* (Cambridge, Mass., Harvard University Press, 1955), pp. 65–68, for parallel remarks concerning the justification of the principles of deductive and inductive inference." *Id.* at 20 n.7.

[54] *See, e.g.*, MIKHAIL, *supra* note 19, at 288-89 (listing nine separate passages in Chapter Six of TAKING RIGHTS SERIOUSLY in which Dworkin refers to reflective equilibrium as a "technique").

[55] RAWLS, *supra* note 1, at 20.

[56] *See, e.g.*, NOAM CHOMSKY, ASPECTS OF A THEORY OF SYNTAX 4 (1965).

[57] RAWLS, *supra* note 1, at 21.

Section 4. For example, he says, "there is no point at which an appeal is made to self-evidence in the traditional sense either of general conceptions or particular convictions" and he elaborates this point as follows:

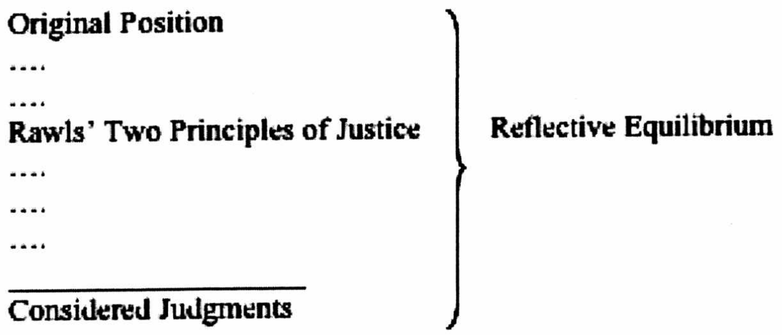
> I do not claim for the principles of justice proposed that they are necessary truths or derivable from such truths. A conception of justice cannot be deduced from self-evident premises or conditions on principles; instead, its justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view.[58]

Rawls' definition of reflective equilibrium in Section 4 can be illustrated by means of the simple diagram in Figure 1 below. According to Rawls' definition, a reflective equilibrium is the state of affairs that is reached when the moral theorist knows the principles to which her chosen set of considered judgments (which themselves may have changed in the process leading to reflective equilibrium) conform, along with the premises of those principles' derivation.[59] When this state of affairs has been reached, the theorist has arrived at a conception of the original position. A considered judgment is justified when it is derivable from the original position. At that point—although not until then—the considered judgment may be called a "considered judgment in reflective equilibrium." Therefore, if we inquire as to the relationship among reflective equilibrium, the justification of considered judgments, and the original position, the answer is that in Rawls' stated framework, these concepts are merely different ways of saying the same thing.

[58] *Id.*

[59] Here I should note that I am taking the liberty of resolving an apparent ambiguity in Rawls' definition of reflective equilibrium. Rawls writes: "It is an equilibrium because at last our principles and judgments coincide; and it is reflective since we know to what principles our judgments conform and the premises of their derivation." *Id.* at 20. What is the antecedent of "their" in the phrase "their derivation"? If the antecedent is "judgments," then Rawls would be saying that a reflective equilibrium is reflective because we know "to what principles our judgments conform and the premises of the derivation of those judgments," which would appear to be redundant. Hence it seems preferable to interpret the intended antecedent of "their" as the principles themselves. On this reading Rawls would be saying that a reflective equilibrium is reflective because we know "to what principles our judgments conform and, in turn, the premises of the derivation of those principles."

Figure 1. Schematic Diagram of Rawls' Definition of Reflective Equilibrium in Section 4 of *A Theory of Justice*



B. The Section 9 Account of Reflective Equilibrium

In Section 9, Rawls characterizes reflective equilibrium in more elaborate terms. The Section 9 definition is more revealing for a number of reasons, perhaps the most significant of which is its precise location in Rawls' text. For it is only after comparing the problem of descriptive adequacy in ethics with the problem of descriptive adequacy in linguistics, and distinguishing moral competence and moral performance by identifying considered judgments as those judgments in which our moral capacities are likely to be displayed without distortion, that Rawls reintroduces the notion of reflective equilibrium in Section 9 that he first discussed in Section 4. "The need for this idea," Rawls now explains, "arises as follows":

> According to the provisional aim of moral philosophy, one might say that justice as fairness is the hypothesis that the principles which would be chosen in the original position are identical with those that match our considered judgments and so these principles describe our sense of justice. But this interpretation is clearly oversimplified. In describing our sense of justice an allowance must be made for the likelihood that considered judgments are no doubt subject to certain irregularities and distortions despite the fact that they are rendered under favorable circumstances. When a person is presented with an intuitively appealing account of

> his sense of justice (one, say, which embodies various reasonable and natural presumptions), he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly. He is especially likely to do this if he can find an explanation for the deviations which undermines his confidence in his original judgments and if the conception presented yields a judgment which he finds he can now accept. From the standpoint of moral philosophy, the best account of a person's sense of justice is not the one which fits his judgment prior to his examining any conception of justice, but rather the one which matches his judgments in reflective equilibrium. As we have seen, this state is one reached after a person has weighed various proposed conceptions and he has either revised his judgments to accord with one of them or held fast to his initial convictions (and the corresponding conception).[60]

On casual glance, this definition of reflective equilibrium seems to repeat the Section 4 definition reviewed earlier. Rawls' use of the phrase "As we have seen" in the last sentence of the paragraph reinforces this impression. However, the account of reflective equilibrium that Rawls provides in this paragraph differs from the Section 4 definition in what seem like crucial respects. In Section 4, reflective equilibrium is a state of affairs that is reached in the course of evaluating proposed interpretations of the initial situation and identifying one interpretation, the original position, as the most favored. By contrast, in this passage reflective equilibrium is a state of affairs that is reached in the course of evaluating competing descriptions of the sense of justice. Likewise, whereas justice as fairness in Section 4 is the contractual thesis that the principles of justice are justified because they are chosen in the original position (thereby qualifying as justifiable on the basis of the notion of pure procedural justice), here justice as fairness is the hypothesis that these same principles describe the sense of justice. In short, whereas in Section 4 Rawls appears to be discussing the *normative* problem of justifying principles of justice, Rawls' topic in Section 9 appears to be the *descriptive* problem of describing and explaining the sense of justice.

[60] *Id.* at 48.

The apparent discrepancy in the definitions of reflective equilibrium that Rawls provides in Sections 4 and 9 raises a number of important questions about the significance of his theory of justice as a whole. Perhaps the most pressing of these questions may be stated as follows. In *A Theory of Justice*, Rawls certainly seems to be defending a solution to the problem of normative adequacy. This seems like the only reasonable interpretation of many statements he makes throughout the book, such as his remarks on the notion of pure procedural justice in Sections 14 and 20 and his remarks on justification in Sections 4 and 87. Yet Rawls also repeatedly describes justice as fairness as a solution to the problem of descriptive adequacy. In Section 9, for example, he repeatedly explains that the goal of moral theory is to produce an accurate description of the sense of justice, which he clearly takes to be a particular cognitive capacity of the human mind:

> Now one may think of moral philosophy . . . as *the attempt to describe our moral capacity*; or, in the present case, one may regard a theory of justice as *describing our sense of justice*.[61]
>
> A conception of justice *characterizes our moral sensibility* when the everyday judgments we do make are in accordance with its principles.[62]
>
> Only a deceptive familiarity with our everyday judgments and our natural readiness to make them could conceal the fact that *characterizing our moral capacities* is an intricate task.[63]
>
> There is no reason to assume that *our sense of justice can be adequately characterized* by familiar common sense precepts, or derived from the more obvious learning principles. *A correct account of moral capacities* will certainly involve principles and theoretical constructions which go much beyond the norms and standards cited in everyday life.[64]

[61] *Id.* at 46 (emphasis added).

[62] *Id.* (emphasis added).

[63] *Id.* at 46–47 (emphasis added).

[64] *Id.* at 47 (emphasis added).

> According to the provisional aim of moral philosophy, one might say that justice as fairness is the hypothesis that the principles which would be chosen in the original position are identical with those that match our considered judgments and so these principles *describe our sense of justice.*[65]
>
> In *describing our sense of justice* an allowance must be made for the likelihood that considered judgments are no doubt subject to certain irregularities and distortions despite the fact that they are rendered under favorable circumstances.[66]
>
> [I]f we can describe one person's sense of grammar we shall surely know many things about the general structure of language. Similarly, if we should be able *to characterize one (educated) person's sense of justice*, we would have a good beginning toward a theory of justice. We may suppose that *everyone has in himself the whole form of a moral conception.*[67]
>
> I wish to stress that a theory of justice is precisely that, namely, a theory. It is a theory of *the moral sentiments* (to recall an eighteenth century title) setting out *the principles governing our moral powers, or, more specifically, our sense of justice.*[68]

All of these characterizations of the fundamental goals of moral theory in Section 9 imply that those goals are primarily descriptive. Moreover, the statements that carry this implication are not limited to Section 9. In Section 20, for example, Rawls explains that the original position "is not intended to explain human conduct except insofar as it tries to account for our moral judgments and helps to explain our having a sense of justice."[69] Rawls then adds: "Justice as fairness is a theory of our moral sentiments as manifested by our considered judgments in reflective equilibrium."[70] Therefore, in light of all these characterizations, it

[65] *Id.* at 48 (emphasis added).
[66] *Id.* (emphasis added).
[67] *Id.* at 50 (emphasis added).
[68] *Id.* at 50-51 (emphasis added).
[69] *Id.* at 120.
[70] *Id.*

seems natural to ask whether Rawls' apparently divergent objectives in *A Theory of Justice*—descriptive and normative—can be reconciled, and if so how, and what role the notion of reflective equilibrium is supposed to play in this process.

C. Resolving the Tension: The Function of Reflective Equilibrium

But are Rawls' aims necessarily inconsistent? Richard Brandt, R.M. Hare, Peter Singer, and other commentators who have criticized Rawls' conception of moral theory for being too empirical and insufficiently normative apparently think so. Nevertheless, to assume that there must be an irresolvable tension between the empirical and normative aspects of Rawls' theory is to miss the point of Rawls' reference to Nelson Goodman in Section 4. The theme of those pages of *Fact, Fiction, and Forecast* to which Rawls directs his reader when he first defines reflective equilibrium in Section 4 is the justification of induction. Goodman's main point in that discussion is that philosophers should be wary of expecting too much from the theory of induction. They should stop plaguing themselves with certain unanswerable questions about justification and should relax the distinction between *justifying* principles of induction and *describing* ordinary, reliable inductive practice. Summarizing this perspective, Goodman writes:

> We no longer demand an explanation for guarantees that we do not have, or seek keys to knowledge that we cannot obtain. It dawns upon us that the traditional smug insistence upon a hard-and-fast line between justifying induction and describing ordinary inductive practice distorts the problem. And we owe belated apologies to Hume. For in dealing with the question how normally accepted inductive judgments are made, he was in fact dealing with the question of inductive validity.[71]

How do Goodman's remarks in this passage bear on Rawls' two definitions of reflective equilibrium in Sections 4 and 9? On the view I wish to propose, they imply that the tension between them is only apparent. There is no necessary inconsistency in

[71] GOODMAN, *supra* note 4, at 64–65.

assuming that a set of moral principles can be part of a solution to the problems of empirical and normative adequacy simultaneously.

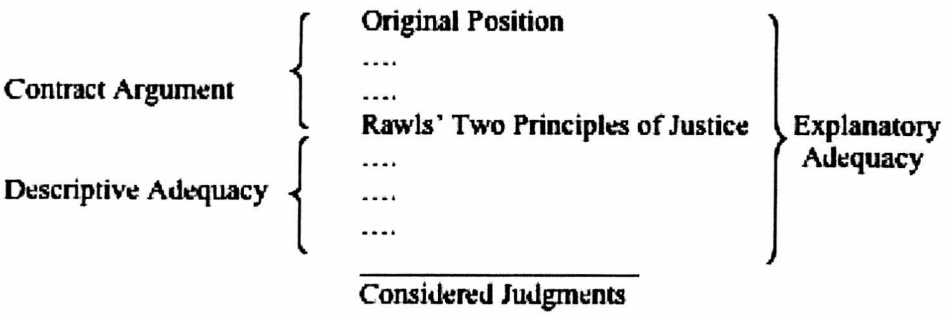
To see this point, it helps to examine how Rawls conceives of the problem of normative adequacy in diagrammatic form (Figure 2 below). In *A Theory of Justice*, Rawls' two principles of justice are what he takes to be a solution to the problem of descriptive adequacy (with respect to one component of human moral competence, namely, the sense of social justice). His contract argument is intended to show that these principles would be chosen in the original position, thereby proving that they are rational. At the same time, Rawls conceives of the original position as supplying a provisional solution to the problem of *explanatory* adequacy—the problem of how moral knowledge or the sense of justice is acquired[72]—insofar as it represents an acquisition model that helps explain the fact that normal persons possess a sense of justice.[73] Hence, when Rawls equates justice as fairness *both* with the claim that his two principles would be chosen in the original position—thereby being part of a solution to the problem of normative adequacy—*and* with the empirical hypothesis that those principles constitute an accurate description and explanation of the sense of justice—thereby solving the problems of descriptive and explanatory adequacy—he is not necessarily being inconsistent. A set of moral principles can, in theory, be descriptive, explanatory, and normative simultaneously. Indeed, according to Rawls, this is the philosophical ideal.[74] It is part of what it means to justify the morality of common sense, by showing that it has a rational foundation.

[72] On the meaning and significance of the problem of explanatory adequacy and its application to moral theory, see generally MIKHAIL, *supra* note 19; Mikhail, *Rawls' Linguistic Analogy*, *supra* note 19.

[73] RAWLS, *supra* note 1, at 120 ("The conception of the original position is not intended to explain human conduct except insofar as it tries to account for our moral judgments and helps to explain our having a sense of justice. Justice as fairness is a theory of our moral sentiments as manifested in our considered judgments in reflective equilibrium.").

[74] "[J]ustice as fairness can be understood as saying that the two principles previously mentioned would be chosen in the original position in preference to other traditional conceptions of justice, for example, those of utility and perfection; and that these principles give a better match with our considered judgments on reflection than these recognized alternatives. Thus justice as fairness moves us closer to the philosophical ideal; it does not, of course, achieve it." *Id.* at 49–50.

Figure 2. Schematic Diagram of Rawls' Account of the Problem of Normative Adequacy in *A Theory of Justice*



It is important to avoid misinterpreting the diagram in Figure 2. It is potentially misleading for several interrelated reasons. In the first place it fails to take into account the theory-dependence of the competence-performance distinction.[75] In Section 9, Rawls emphasizes that moral theorists may want to change what they presently take to be considered judgments once those judgments' regulative principles are brought to light. So it is important not to be misled by the static appearance of the diagram in Figure 2. A reflective equilibrium is not necessarily stable. As its name implies, it is always subject to further reflection and change.[76]

Second, it is important not to be misled by Rawls' reference to Goodman and the problem of justifying induction. It is tempting to conclude that just as "[t]he problem of justifying induction is not a problem of demonstration but a problem of defining the difference between valid and invalid inductions,"[77] so too, according to Rawls, is the problem of justifying moral principles not something over and above the problem of describing or defining what one takes, pre-theoretically, to be instances of valid moral reasoning. This interpretation of Rawls, however, runs the risk of ignoring an important aspect of how he conceives of the problem of normative adequacy in ethics, which has no clear analogue in the case of either induction or linguistics—namely, the requirement of rationality, understood to be the suitability of moral principles as

[75] On the meaning and significance of the competence-performance distinction in both linguistics and moral theory, see generally MIKHAIL, *supra* note 19; Mikhail, *Rawls' Linguistic Analogy*, *supra* note19.

[76] MIKHAIL, *supra* note 19, at 20–21; *cf.* Nagel, *supra* note 7, at 2–3.

[77] GOODMAN, *supra* note 4, at 65.

the object of rational choice. Neither induction nor linguistics requires the principles of an adequate theory to be rational in this special sense.

IV. THE DISTINCTION BETWEEN NARROW AND WIDE REFLECTIVE EQUILIBRIUM

Figure 2 is an oversimplification for a third reason: it fails to allow for the possibility that when a person is presented with an intuitively appealing account of her sense of justice, her sense of justice may itself be transformed as a result of reflecting on this account and considering its implications. This is the basis for the distinction that Rawls draws between *narrow* and *wide* reflective equilibrium, which Rawls first explains in the following terms in Section 9:

> There are . . . several interpretations of reflective equilibrium. For the notion varies depending upon whether one is to be presented with only those descriptions which more or less match one's existing judgments except for minor discrepancies, or whether one is to be presented with all possible descriptions to which one might plausibly conform one's judgments together with all relevant philosophical arguments for them. In the first case we would be describing a person's sense of justice more or less as it is although allowing for the smoothing out of certain irregularities; in the second case a person's sense of justice may or may not undergo a radical shift. Clearly it is the second kind of reflective equilibrium that one is concerned with in moral philosophy.[78]

In this passage, Rawls uses the phrase "second kind of reflective equilibrium" to refer to what he calls "wide reflective equilibrium" in *Independence*. By invoking this second kind of reflective equilibrium, Rawls makes a reasonable allowance for the possibility that when a person is given the opportunity to reflect on an empirically adequate theory of her sense of justice, her sense of justice may undergo a dramatic shift. Although he does not elaborate on this observation to a great extent, its meaning and

[78] RAWLS, *supra* note 1, at 49.

motivation seem fairly clear. Rawls' overriding aim in *A Theory of Justice* is not merely to convince his reader that justice as fairness is a better overall account of the human sense of justice than utilitarianism. He clearly wants to do this, thereby convincing those readers previously inclined toward utilitarianism that their prior understanding of their own sense of justice may be mistaken. Yet he also wishes to leave open the possibility that the act of reflecting on a moral theory may cause a person's sense of justice to be transformed. Hence, he allows for the possibility that reading a book like *A Theory of Justice* will cause a person whose sense of justice is utilitarian to change her mind.

In *Independence*, Rawls revisits the distinction between narrow and wide reflective equilibrium in an important passage that uses the terms "narrow" and "wide" for the first time. First, he emphasizes that in attempting to construct an adequate theory of the human moral sense or sense of justice, a researcher must be careful to distinguish her standpoint as a moral theorist from her standpoint as a person applying her own moral beliefs to particular moral and social problems:

> In order to [investigate the moral conceptions that people hold, or would hold under suitably defined conditions], one tries to find a scheme of principles that match people's considered judgments and general convictions in reflective equilibrium. This scheme of principles represents their moral conception and characterizes their moral sensibility. One thinks of the moral theorist as an observer, so to speak, who seeks to set out the structure of other people's moral conceptions and attitudes. Because it seems likely that people hold different conceptions, and the structure of these conceptions is in any case hard to delineate, we can best proceed by studying the main conceptions found in the tradition of moral philosophy and in leading representative writers, including their discussions of particular moral and social issues. We may also include ourselves, since we are ready to hand for detailed self-examination. But in studying oneself, one must separate one's role as a moral theorist from one's role as someone who has a particular conception. In the former role we are investigating an aspect of human psychology, the structure of our moral sensibility; in the latter we

> are applying a moral conception, which we may regard (though not necessarily) as a correct theory about what is objectively right and wrong.[79]

Two paragraphs later, Rawls reiterates that he conceives of the moral theorist as an observer whose aim is to characterize the implicit moral beliefs of people in general. He explains the meaning of the distinction between narrow and wide reflective equilibrium in this context:

> Furthermore, because our inquiry is philosophically motivated, we are interested in what conceptions people would affirm when they have achieved wide and not just narrow reflective equilibrium, an equilibrium that satisfies certain conditions of rationality. That is, adopting the role of observing moral theorists, we investigate what principles people would acknowledge and accept the consequences of when they have had the opportunity to consider other plausible conceptions and to assess their supporting grounds. Taking this process to the limit, one seeks the conception, or plurality of conceptions, that would survive the rational consideration of all feasible conceptions and all reasonable arguments for them. We cannot, of course, actually do this, but we can do what seems like the next best thing, namely, to characterize the structures of the predominant conceptions familiar to us from the philosophical tradition, and to work out the further refinements of these that strike us as most promising.[80]

The account of the distinction between narrow and wide reflective equilibrium given in this passage is virtually identical to Rawls' initial account of the same distinction in Section 9 of *A Theory of Justice*. Indeed, there are only two discernible differences between the two accounts. First, "considering all possible descriptions to which one might plausibly conform one's judgments" in Section 9 becomes considering "other plausible conceptions" in *Independence*. Second, considering conceptions of justice "together with all relevant philosophical arguments for

[79] Rawls, *supra* note 9, at 7.

[80] *Id.* at 8.

them" in Section 9 becomes conducting a "philosophically motivated" inquiry and searching for an equilibrium that "satisfies certain conditions of rationality" by assessing the "supporting grounds" of different moral conceptions in *Independence*. Otherwise the two explanations are identical. Thus, we can summarize Rawls' original distinction between narrow and wide reflective equilibrium as follows. Rawls takes a narrow reflective equilibrium to be the state of affairs in which a description of the sense of justice that has been selected from a relatively narrow set of possible descriptions more or less matches an existing set of considered judgments. By contrast, a wide reflective equilibrium exists when such a description has been selected from a much wider range of alternatives and the individuals in question have had the opportunity to consider all relevant philosophical arguments in support of those alternatives, whereupon their sense of justice may or may not have undergone a radical shift.

V. DANIELS' REINTERPRETATION OF NARROW AND WIDE REFLECTIVE EQUILIBRIUM

Carefully identifying the differences between the two explanations Rawls gives of the distinction between narrow and wide reflective equilibrium in *A Theory of Justice* and *Independence* may seem pedantic or unnecessary. Doing so is important, however, because it allows us to formulate a significant objection to one of the most influential interpretations of both reflective equilibrium and Rawls' linguistic analogy in the philosophical literature: namely, that of Norman Daniels.[81] Daniels' stated objective in his paper, *On Some Methods of Ethics and Linguistics*, is "to free wide equilibrium from an unnecessary or, at least, overstated analogy to linguistic method."[82] Yet in a footnote,[83] it becomes apparent that Daniels does not seek to "free" moral theory from comparisons to what linguists refer to as E-language (i.e., externalized) or P-language (i.e., Platonic) interpretations of linguistics,[84] but only from what linguists refer to as I-language (i.e., internalized) interpretations, such as

[81] *See generally* DANIELS, *supra* note 10.

[82] *Id.* at 66.

[83] *Id.* at 79 n.17.

[84] For a discussion of P-language (i.e., Platonic) interpretations of linguistics, see generally JERROLD KATZ, LANGUAGE AND OTHER ABSTRACT OBJECTS (1981). *See also* ROBERT NOZICK, PHILOSOPHICAL EXPLANATIONS 744 (1981) (discussing Katz's Platonist interpretation).

Chomsky's, in which the subject matter of linguistic theory is an aspect of human psychology—a distinctive language faculty that exists "inside the head." Thus, in his footnote Daniels explains that his criticism of the linguistic analogy is "restricted to approaches to syntactics that view it as a branch of psychology, broadly construed. These are the approaches Rawls has in mind in proposing the analogy, but alternative approaches, such as the one indicated in Jerrold Katz's recent work, reject the psychologizing of linguistics."[85]

Daniels' main claim in his paper is that it is "wide equilibrium . . . not narrow, that is of interest to the moral philosopher—and for just those reasons that distinguish it from syntactics."[86] Significantly, however, Daniels characterizes these two forms of reflective equilibrium differently than Rawls does. Rawls does not take the difference between wide and narrow reflective equilibrium to map onto the distinction between an I-morality conception of moral theory and some other conception, according to which moral theory is not conceived as part of psychology, broadly construed. On the contrary, Rawls considers both narrow and wide reflective equilibrium to be states of affairs that are achieved in the course of "investigating an aspect of human psychology, the structure of our moral sensibility."[87] By contrast, although Daniels' definitions are abstract and faithful to Rawls' own accounts as far as they go—defining narrow reflective equilibrium as "an ordered pair of (a) a set of considered moral judgments acceptable to a given person *P* at a given time, and (b) a set of general moral principles that economically systematizes (a),"[88] and wide reflective equilibrium as an ordered triple of (a), (b), and "(c) the set of relevant theories invoked or presupposed by the winning arguments for (b), all duly 'adjusted' to be compatible with each other,"[89]—he also takes the further unwarranted step of effectively stipulating that the target of wide reflective equilibrium is *not* the human moral sense or sense of justice, or indeed any other aspect of human psychology.[90] What Daniels actually produces in his paper, therefore, is neither a coherent criticism of the linguistic analogy, nor a coherent criticism of the conception of moral theory Rawls actually describes in *A Theory of Justice*, but a terminological sleight of hand. Daniels simply redefines narrow and wide reflective

[85] DANIELS, *supra* note 10, at 79 n.17.
[86] *Id.* at 66.
[87] Rawls, *supra* note 9, at 7.
[88] DANIELS, *supra* note 10, at 67.
[89] *Id.* at 70.
[90] *Id.*

equilibrium in a manner that dissociates the latter from an I-morality interpretation of moral theory. He then draws on this new definition to reject the linguistic analogy, first by arguing that wide reflective equilibrium (as he defines it) is at odds with the scientific methods of linguistics in various respects, and then by concluding, on the foregoing basis, that "the heart of the analogy to the case of descriptive syntactics is gone."[91] The main effect of all this theoretical maneuvering is to obscure a substantive difference between Daniels and Rawls over a fundamental issue, namely, the proper subject matter of moral theory and the significance of a descriptively adequate moral psychology with respect to it.

Of course Daniels is free to define or redefine technical terminology in any way he wants. He is also at liberty to disagree with Rawls that the proper subject matter of moral theory is an aspect of human psychology. But one would think that he owes it to his readers to make clear that this is what he is doing. Instead of this, Daniels simply assumes that descriptive adequacy "is not of central interest to moral philosophy"[92]—essentially the same strategy of anti-psychologism and avoidance that one finds in Hare and Singer—and thereby constructs a mere pseudo-argument against the linguistic analogy and its significance for moral philosophy. Moreoever, just like Hare and Singer, Daniels frequently begs the very questions about the nature and origin of human moral intuitions that a research program in moral theory inspired by the linguistic analogy is meant to answer in the first place.[93]

VI. CONCLUSION

The distinction between the problems of descriptive and normative adequacy that motivates and informs Rawls' concept of reflective equilibrium is a recurring feature of the conception of moral theory that can be found in his early writings, including *Grounds*, *Outline*, *A Theory of Justice*, and *Independence*. In *Grounds* and *Outline*, Rawls adopts the view that there is an order of priority between the problem of descriptive adequacy and the problem of normative adequacy, the descriptive taking precedence over the normative. He also assumes that a solution to the problem of descriptive adequacy constitutes a presumptive solution to the

[91] *Id.* at 72.
[92] *Id.* at 69.
[93] *See, e.g.*, *id.* at 70–72.

problem of normative adequacy, given the kind of evidence that a descriptively adequate set of principles explains.[94] In *A Theory of Justice*, Rawls' approach to the problem of metaethical adequacy—the difficult philosophical question of *how* moral principles can be justified—is more complex. Specifically, as I interpret him, Rawls makes at least five major modifications to his approach to metaethical adequacy in *A Theory of Justice*. First, he includes the question of how moral knowledge or a sense of justice is acquired to be among the fundamental problems that he thinks a comprehensive moral theory should solve.[95] Second, he introduces the contract argument and the notion of pure procedural justice as means of showing that the principles of justice as fairness are rational. Third, he follows Goodman in defining reflective equilibrium in a manner that allows for solutions to the problems of empirical and normative adequacy to be mutually dependent. Fourth, he makes certain reasonable allowances for the theory-dependence of the competence–performance distinction, allowing that the set of considered judgments can change over time in light of empirical inquiry and philosophical reflection. Finally, Rawls adopts the weakest of three possible metaethical standpoints corresponding to the familiar three-fold distinction among discovery, decision, and evaluation procedures that Chomsky identifies in *Syntactic Structures*,[96] holding that the overriding goal of a moral theory is to construct an evaluation procedure for moral principles, rather than seeking to accomplish any more ambitious theoretical goal. Consequently Rawls remains satisfied with the relatively modest claim that justice as fairness is a *better* overall account of the sense of justice than either utilitarianism or its other rivals. As Rawls puts it, his main claim in this regard is that justice as fairness is "more reasonable than" any of its rivals, hence it is "justifiable with respect to [them]."[97]

Brandt, Hare, Singer, and other critics who have objected to Rawls' early conception of moral theory on the grounds that it is too empirical and insufficiently normative have given no indication

[94] *See generally* MIKHAIL, *supra* note 19.

[95] Rawls' concern with the acquisition problem first emerges most clearly in an important article published in 1963. *See* John Rawls, *The Sense of Justice*, 72 PHIL. REV. 281 (1963). His reflections on this problem are elaborated at length in the third part of *A Theory of Justice*, particularly Chapter Eight. *See* RAWLS, *supra* note 1, at 453–512.

[96] On the meaning and significance of Chomsky's three-fold distinction among discovery procedures, decision procedures, and evaluation procedures and its application to moral theory, see generally Mikhail, *Rawls' Linguistic Analogy*, *supra* note 19.

[97] RAWLS, *supra* note 1, at 17.

in their published writings that they have grasped the full complexity of Rawls' approach to the problem of metaethical adequacy. Nor, so far as I am aware, have they adequately defended a more cogent alternative. For his part, Daniels does not defend Rawls' original concept of reflective equilibrium in *A Theory of Justice* as much as reinterpret it in order to change the topic, attempting to shift philosophers' attention away from moral psychology in the unconvincing manner that has become all-too-familiar in recent years.[98] Rawls' original concept and its implications for the complex interaction among descriptive moral psychology, normative ethics, and metaethics is subtle, powerful, and compelling, and unless and until a more cogent alternative is actually produced and shown to be superior to it, his account of reflective equilibrium and of the complex interdependence of these different branches of ethics in *A Theory of Justice* would appear to be the most defensible account available in the literature. It takes a theory to beat a theory.[99] As is the case with any scientific or philosophical theory, the best a moral theory can hope to achieve is to be better than its alternatives.

[98] *See generally* MIKHAIL, *supra* note 19.

[99] *See* Richard A. Epstein, *Common Law, Labor Law, and Reality: A Rejoinder to Professors Getman and Kohler*, 92 YALE L. J. 1435, 1435 (1983) (offering the first known use of this familiar phrase in the legal literature). *But cf.* Thomas S. Ulen, *A Nobel Prize in Legal Science: Theory, Empirical Work, and the Scientific Method in the Study of Law*, 2002 U. ILL. L. REV. 875, 887 n. 47 (2002) (questioning whether "it takes a theory to beat a theory" accurately describes how scientific advances typically occur). For both of these citations, I am indebted to Professor Larry Solum, and in particular to the entry on "It Takes a Theory to Beat a Theory" in Professor Solum's *Legal Theory Lexicon*.