Michael Baur

Reversing Rawls

Criteriology, contractualism and the primacy of the practical

Abstract In this paper, I offer an immanent critique of John Rawls's theory of justice which seeks to show that Rawls's understanding of his theory of justice as *criteriological* and *contractarian* is ultimately incompatible with his claim that the theory is grounded on *the primacy of the practical*. I agree with Michael Sandel's observation that the Rawlsian theory of justice rests on substantive metaphysical and epistemological claims, in spite of Rawls's assurances to the contrary. But while Sandel argues for even *more* substantive metaphysical and epistemological commitments, I argue in the opposite direction. Following J. G. Fichte, I argue for a normative theory of society, not based on some particular notion of the good or on some contentious account of what all reasonable persons would agree to, but based only on the *radical* primacy of the practical, that is, based only on the seemingly empty premise that free beings – precisely because they are free – cannot be imagined in advance as all agreeing to any particular thing at all.

Key words contractualism · J. G. Fichte · primacy of the practical · Rawls

In this paper, I offer an immanent critique of John Rawls's theory of justice, that is, a critique whose force depends only on the uncovering of difficulties internal to the theory itself, and not on claims drawn from sources external to it. More specifically, I seek to show that Rawls's understanding of his theory of justice as *criteriological* and *contractarian* is ultimately incompatible with his claim that the theory is grounded on *the primacy of the practical* (i.e. that it is merely political, not metaphysical or epistemological).[1] At the end of this paper, I shall indicate briefly how my immanent critique of Rawls's theory of justice might point the way towards a more adequate normative theory of society,

PHILOSOPHY & SOCIAL CRITICISM • **vol 28 no 3** • pp. 251–296 PSC

[0191-4537(200205)28:3;251–296;022811]

one suggested by Kant's famous – but often misunderstood – younger contemporary, Johann Gottlieb Fichte.[2] However, since my critique is an *immanent* one, I shall reserve all references to Fichte for the very end of this paper.

As will become clear later, I generally agree with Michael Sandel's observation that the Rawlsian theory of justice rests on substantive metaphysical and epistemological claims, in spite of Rawls's assurances to the contrary.[3] But unlike Sandel, I argue that a genuinely immanent critique should not seek to show that Rawls should have made such substantive (metaphysical and epistemological) claims more forthrightly and robustly[4] – for even if the theory of justice does rely on such claims, Rawls would still want to argue that such reliance is undesirable, since it is incompatible with 'the primacy of the practical'. Thus while Sandel seeks to correct Rawls by calling for even *more* substantive metaphysical and epistemological commitments, I argue that the Rawlsian project can and should be made more consistent in the opposite direction, through a greater emphasis on the primacy of the practical. And so, following Fichte (and not Rawls or Sandel), I argue for a normative theory of society, based not on some particular notion of the good or on some contentious account of what all reasonable persons would agree to, but based only on the *radical* primacy of the practical, that is, based only on the seemingly empty premise that free beings – precisely because they are free – cannot be imagined in advance as all agreeing to any particular thing at all.

My argument as a whole will unfold in six parts:

Part I explains why the Rawlsian project is to be understood as a *criteriological* project.

Part II aims to show why Rawls's criteriological project is also a *political* and *social contractarian* project, according to which the source of the criterion being sought (namely, the principles of justice) is nothing other than personhood insofar as it is said to be *self-determining* in the relevant respects.

Part III explains why the personhood that is the source of the Rawlsian criterion must be *conceptually different* from our own personhood, and thus why Rawls must distinguish between two types of personhood (namely, our personhood and personhood in the original position).

Part IV explains how Rawls's separation of personhood into two types creates internal difficulties for his account of how the principles of justice are to be *derived*.

Part V explains how Rawls's separation of personhood into two types creates internal difficulties for his account of how the principles of justice are to be *applied*.

Part VI seeks to show how Fichte can be understood as offering a more plausible, and more consistent, theory of justice, one that is grounded on a radicalized conception of freedom and the primacy of the practical.

I believe that my immanent critique of Rawls's theory of justice applies equally well to any other contractarian theory of society that – like Rawls's – would seek to ground the principles of justice on the separation of personhood into two types (our personhood and personhood in an ideal agreement situation); however, my limited aim in this paper is simply to articulate the immanent critique insofar as it applies to Rawls's theory and on Rawls's own terms.

I Rawls's project as a criteriological project

Basic social and legal institutions can be configured in a wide variety of ways. There is no particular pattern or single arrangement that ineluctably imposes itself on our modes of production and interaction. Accordingly, it makes sense to ask and deliberate about *normative* (and not merely descriptive) questions pertaining to our basic social and legal institutions: e.g. questions about how these institutions *ought* to be individually designed and how they *ought* to be combined in a general pattern of social interaction (regardless of how they are actually designed and combined). According to John Rawls, a reasoned answer to these kinds of normative questions about social arrangements must refer to some criterion or standard 'for choosing among . . . various social arrangements'.[5] The problem of assessing and choosing from among various possible social arrangements can thus be understood as a *criteriological* problem, i.e. a problem of identifying the right criterion or standard for assessing and choosing from among various possible social arrangements.

But what is meant by 'society' or 'social arrangements'? According to Rawls, society can broadly be understood as a 'cooperative venture for mutual advantage':

> Let us assume, to fix ideas, that a society is a more or less self-sufficient association of persons who in their relations to one another recognize certain rules of conduct as binding and who for the most part act in accordance with them. Suppose further that these rules specify a system of cooperation designed to advance the good of those taking part in it. Then, although society is a cooperative venture for mutual advantage, it is typically marked by a conflict as well as by an identity of interests.[6]

There is an identity of interests, because 'social cooperation makes possible a better life for all than any would have if each were to live by

his own efforts'.[7] However, there is also a conflict of interests, since 'persons are not indifferent as to how the greater benefits produced by their collaboration are distributed, for in order to pursue their ends they each prefer a larger to a lesser share'.[8] For Rawls, then, society is characterized by both agreement and division.

Furthermore, Rawls argues that we need a criterion or standard for the purpose of assessing and choosing from among the ways that society, as a unity-in-plurality of interests, might be organized. Since the requisite criterion would guide our judgements about the distribution of benefits and burdens among different persons within society, the criterion would amount to a set of *principles of justice*:

> A set of principles is required for choosing among the various social arrangements which determine this division of advantages and for underwriting an agreement on the proper distributive shares. These principles are the principles of social justice: they provide a way of assigning rights and duties in the basic institutions of society and they define the appropriate distribution of the benefits and burdens of social cooperation.[9]

It is worth noting that we (and Rawls) have spoken about a 'criterion' or 'standard' in the singular, but of the 'principles of justice' in the plural. This linguistic difference, of course, points to a potential problem for Rawls's entire criteriological project. If the proposed criterion or standard were a set of *several* (i.e. two or more) non-ordered principles, then it could not really perform its criteriological function in guiding our judgements. The reason for this is not too difficult to grasp. Wherever there is not just one principle of justice but a set of several, non-ordered principles, it is possible that different orderings of the same set of principles might lead to different normative judgements concerning the justness or unjustness of a particular social arrangement. That is, the very same set of principles might be able to support contradictory normative judgements; and any choice between the two (or more) conflicting normative judgements (or the different orderings of the principles) would be *ad hoc* or arbitrary from the point of view of the principles themselves. If that were the case, then the *several* principles would not really be serving their criteriological purpose. In order to arrive at a set of several principles that can serve as a valid criterion or standard, it is necessary *to reduce* the several principles to a unity (in which there would not be several principles, but only one), or to impose *a supervening unity* on the several principles (by ranking them in order of importance or priority).

Now Rawls is fully aware that an unreduced or non-ordered plurality of principles would be fatal to the criteriological project that he is pursuing. Accordingly, he argues that the principles that he will

eventually propose are to be ranked in serial or 'lexical' order.[10] In effect, the lexical ordering of several principles introduces a supervening unity on the principles. This is because a lexical order is one

> . . . which requires us to satisfy the first principle in the ordering before we can move on to the second, the second before we consider the third, and so on. A principle does not come into play until those previous to it are either fully met or do not apply. A serial ordering avoids, then, having to balance principles at all; those earlier in the ordering have an absolute weight, so to speak, with respect to later ones, and hold without exception.[11]

A set of several non-ordered principles cannot serve as a valid criterion or standard for making normative judgements about social arrangements. Rawls's lexical ordering of the principles of justice overcomes a problem that otherwise would be fatal to his criteriological project.

Rawls's concern about ranking the principles of justice in lexical order is part of his effort to avoid 'intuitionism', the doctrine that 'there is an irreducible family of first principles which have to be weighed against one another by asking ourselves which balance, in our considered judgement, is the most just'.[12] According to the intuitionist, there is 'no single standard' that accounts for or assigns different weights to the several ethical principles upon which we base our normative judgements. In order to derive any particular normative judgement from the given set of irreducible, non-ordered principles, we must simply 'strike a balance by intuition'.[13] Applied to the problem of social justice, intuitionism

> . . . holds that in our judgements of social justice we must eventually reach a plurality of first principles in regard to which we can only say that it seems to us more correct to balance them this way rather than that.[14]

Now in order to pursue his criteriological project, Rawls clearly needs to avoid the intuitionist conclusion. For if the intuitionist claim were to hold sway, then our putative reliance on a set of principles as a standard or criterion would necessarily give way to *ad hoc* balancing 'unguided by constructive and recognizably ethical criteria'.[15]

Rawls's overall criteriological purpose in *A Theory of Justice* is indicated by his now famous comparison between justice and truth: 'Justice is the first virtue of social institutions, as truth is of systems of thought.'[16] Just as an account of the principles of 'truth' can serve us in determining what counts as a 'true' system of thought, so too an account of the principles of justice can serve us in determining what counts as a 'just' social arrangement. Just as a conception of truth can be regarded as an epistemological standard, so too '[a] conception of social justice . . . is to be regarded as providing in the first instance a

standard whereby the distributive aspects of the basic structure of society are to be assessed'.[17] The criterion or standard being sought, of course, can be both retrospective and prospective; it can serve both a critical and a prescriptive function. The principles of social justice are to serve as a criterion or standard for criticizing social institutions that *actually do exist* and for designing social institutions that perhaps *ought to exist*. Furthermore, the principles of justice, once articulated, are to exhibit some degree of fixity and stability. After all, they cannot function as a genuinely critical and prescriptive standard if their content is constantly changing throughout their various possible applications. Thus, once articulated, the principles of justice are supposed to 'regulate *all* subsequent criticism and reform of institutions'.[18] Rawls goes so far as to say that the principles of justice implicitly contain 'an ideal of the person that provides *an Archimedean point* for judging the basic structure of society'.[19]

Before going further, it is appropriate to make two brief disclaimers on Rawls's behalf. First, although Rawls describes the content of the principles of justice by reference to the notion of an 'Archimedean point', it would be wrong to think that his derivation of these principles is hopelessly aprioristic, rationalistic, or insensitive to human situatedness and contingency. For Rawls, the process of thought that eventually leads to the articulation of the principles of justice is to be understood as the result of a 'back and forth' movement in thought between *general principles of justice* and *our existing considered convictions*. If our considered convictions do not match up with the general principles, then we can revise the latter in light of the former, or vice versa. Eventually, we should be able to reach what Rawls calls a 'reflective equilibrium', a state in which 'at last our principles and judgements coincide'.[20] Such reflective equilibrium 'is not necessarily stable', and it is 'liable to be upset by further examination'.[21] Yet our arrival at such a reflective equilibrium provides us with a non-apriori, non-rationalistic starting-point for deriving the principles of justice.[22]

Secondly, while Rawls aims to articulate principles that will guide our normative judgements about justice, his purpose is limited in scope. The principles are not to serve as a criterion for making judgements about *anything* that might be said to be ' just' or 'unjust', e.g. acts, events, or persons. Rather, they are to guide us in assessing, criticizing, and/or reforming 'the basic structure of society', i.e. 'the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation'.[23] Since, for Rawls, 'society' is a 'more or less self-sufficient association of persons',[24] the 'basic structure' refers only to those institutions that together constitute a more or less *self-contained* system, namely 'the political constitution and the principal economic and social arrangements'.[25]

II Self-determining personhood as the immediate source or ground of the principles of justice

Thus far we have spoken only very generally about the criteriological purpose, the non-intuitionistic lexical ordering, the non-rationalistic method of derivation, and the limited scope of the principles of justice. Now we have to ask about *the normative source or ground* of the principles of justice. As we have already seen, the lexically ordered principles of justice are to serve as the criterion or standard for our normative judgements about actual or potential social arrangements. But on the basis of what further ground or criterion do we derive and/or justify these principles? In raising this question, one is implicitly raising the problem of a possible infinite regress: if our normative judgements about social arrangements are to be based on the principles of justice, and if these principles themselves are to be based on some other source, then one can quite reasonably ask whether this other source requires yet another source or ground, and so on *ad infinitum*. Given the *criteriological* purpose of Rawls's project, this question is not out of place; for the question of 'the criterion' is essentially a question about first principles; and it is perfectly reasonable to ask about whether one's first principles are genuinely *first* principles, or merely intermediate principles that need to be grounded on the basis of some other, 'truly first' principles. Along these lines, some commentators have expressed surprise and concern that Rawls has not provided a more comprehensive, epistemological justification of his derivation of the principles of justice.[26]

Now, for Rawls there is no danger of a problematic infinite regress, and there is no need for an explicit epistemological justification of the principles of justice, since the immediate source or ground of the principles of justice needs no ground outside of itself, but rather is *self-determining or self-grounding in the relevant respects*. This is because the immediate source or ground of the principles of justice is nothing other than the very personhood or selfhood whose relative rights and duties are at issue; and thus the immediate source or ground is this personhood simply insofar as it chooses *its own* principles for *its own* (self-determined) practical, political purposes. Since this personhood aims to articulate principles of justice in order to evaluate society as a 'cooperative venture for mutual advantage', and since what counts as 'mutual advantage' is to be determined by the persons themselves, the immediate source or ground of the principles of justice, for Rawls, can be nothing other than personhood or selfhood insofar as it chooses principles of justice *for itself*. For Rawls, then, there is a self-referentiality, and thus an implicit self-justification, built into the personhood that is the immediate source or ground of the principles of justice: 'Just as each

person must decide by rational reflection what constitutes his good, that is, the system of ends which it is rational for him to pursue, so a group of persons must decide once and for all what is to count among them as just and unjust.'[27]

The basic point here merits further reflection. Rawls's criteriological project in *A Theory of Justice* is to articulate a standard or criterion for assessing society as a unity-in-plurality of interests, as a cooperative venture for mutual advantage. Now for Rawls, what counts as 'mutual advantage' is not to be decided by reference to any independently existing standard (such as 'nature' or 'God' or a teleological order of things), i.e. a standard that is allegedly antecedent to persons' *own* purposes in organizing themselves into a political association for mutual advantage. Rather, it is to be determined by nothing other than practical, deliberative personhood insofar as it decides *for itself* what shall constitute the terms of fair cooperation for mutual advantage. There is no problem of an infinite regress here, since Rawls's criteriological project does not require any ground or foundation that is prior to, external to, or independent of persons' own self-determined decisions in organizing themselves into a self-sufficient social system.

Another way to express this is to say that the personhood that is the immediate source of the principles of justice is not and need not be interested in any independent epistemological, metaphysical, or theoretical ground for its derivation of the principles of justice; it is interested only in a *practical* or *political* one. Accordingly, such personhood does not and need not refer to any epistemological, metaphysical, or theoretical ground that might antecedently or independently determine for it what its 'true' ends or 'true' nature should be. Instead, such personhood depends only on its own practical and political considerations in organizing itself into a cooperative venture for mutual advantage. Now, of course, *if* the Rawlsian derivation of the principles of justice were to rely on some kind of epistemological, metaphysical, or theoretical ground that were independent of or antecedent to the self-determination of personhood itself, then it *would* give rise to the problem of a possible infinite regress, as noted above. For such an independent epistemological, metaphysical, or theoretical ground would have to be justified by reference to some source *external* to personhood (e.g. nature or God or a teleological order of things); and, in turn, this external source (assuming again that it is not self-determining personhood itself) would have to be justified by reference to yet another external source or ground, and so on *ad infinitum*. In sum: while the Rawlsian project is a criteriological project, it claims to require no external epistemological, metaphysical, or theoretical criterion (and thus is not in danger of engendering an infinite regress), because it is essentially a *practical* or *political* criteriological project, undertaken for the purpose of assessing

the self-organization of personhood itself; and what counts for the purpose of such self-organization is determined by the very personhood that is organized and doing the organizing.

Now the reason why the Rawlsian project is a *practical* and *political* criteriological project (and not an epistemological or metaphysical one) is also the reason why it is a *social contractarian* project. Like other social contractarians before him, Rawls begins with the intuitively appealing idea that the principles of justice are to have their ultimate source in personhood or selfhood insofar as such personhood or selfhood chooses principles for itself and for its own purpose, which is the purpose of organizing itself into a cooperative venture for mutual advantage.[28] Thus for the social contractarian, the criteriological question concerning the principles of justice is essentially a *practical* and *political* question of acceptability or consent, and not one of metaphysics, epistemology, or theoretical standards that are given independently of or antecedently to self-determining selfhood:

> . . . the guiding idea is that the principles of justice for the basic structure of society are the object of the original agreement. They are the principles that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association.[29]

The initial choice situation (within which self-determining personhood chooses principles for its own practical, political self-organization) is called the 'original position' by Rawls. Before saying more about the original position and about the personhood that is operative in the original position, it is necessary first to say something about some apparent differences between Rawls's earlier thought (e.g. as reflected in *A Theory of Justice*) and his later thought (e.g. as reflected in *Political Liberalism*).

It is generally recognized that Rawls – in his work that post-dates *A Theory of Justice*, first published in 1971 – becomes increasingly clear about the practical, political, non-epistemological and self-justificatory nature of his criteriological project.[30] For example, in an article from 1980 (nine years after the publication of *A Theory of Justice* and thirteen years prior to the publication of *Political Liberalism*), Rawls is very clear about the practical and political (i.e. non-metaphysical and non-epistemological) nature of his criteriological project: 'What justifies a conception of justice is not its being true to an order antecedent to and given to us, but its congruence with our deeper understanding of ourselves and our aspirations.'[31] In that same article, Rawls explicitly describes his criteriological project in terms of the 'primacy of the practical':

> The search for reasonable grounds for reaching agreement rooted in our conception of ourselves and in our relation to society replaces the search

> for moral truth interpreted as fixed by a prior and independent order of objects and relations, whether natural or divine, an order apart and distinct from how we conceive of ourselves. The task is to articulate a public conception of justice that all can live with who regard their person and their relation to society in a certain way. And though doing this may involve settling theoretical difficulties, the practical social task is primary.[32]

Now in *Political Liberalism* (published in 1993), Rawls continues to emphasize the primacy of the practical, and the need to understand the personhood that is the source of the principles of justice without reference to any antecedent order of things or to any theory of nature, God, or teleology. Rawls refers to this as the 'political' (as opposed to an epistemological, metaphysical, or teleological) conception of the person.[33] Also in *Political Liberalism*, Rawls criticizes *A Theory of Justice* for having relied on 'a comprehensive philosophical doctrine'. Such a doctrine, according to Rawls, contradicts or undermines the political (not metaphysical or teleological) purpose of the project of articulating the principles of justice. This is because, for Rawls, a comprehensive doctrine includes claims and commitments that are non-political, and therefore extraneous to a properly political criteriological project. According to Rawls's definition, a doctrine is comprehensive 'when it includes conceptions of what is of value in human life, as well as ideals of personal virtue and character, that are to inform much of our non-political conduct (in the limit of our life as a whole)'.[34] Now Rawls also distinguishes between doctrines that are 'fully comprehensive' and those that are only 'partially comprehensive': 'A doctrine is fully comprehensive when it covers all recognized values and virtues within one rather precisely articulated scheme of thought; whereas a doctrine is only partially comprehensive when it comprises certain (but not all) non-political values and virtues and is rather loosely articulated.'[35] But regardless of whether a doctrine is fully or partially comprehensive, any such doctrine – to the extent that it presupposes any purpose or order that is independent of or extraneous to the project of 'social cooperation for mutual advantage' – runs afoul of a properly *political* criteriological project. As Rawls notes, 'for a conception to be even partially comprehensive, it must extend beyond the political and include non-political values and virtues'.[36] To the extent that *A Theory of Justice* does rely on any (fully or partially) comprehensive doctrine, it runs afoul of the purely political and self-justificatory nature of Rawls's own criteriological project, and therefore must be reformed.

As noted above, Rawls criticizes *A Theory of Justice* for having contained a 'comprehensive philosophical doctrine', namely a Kantian one. This self-criticism, of course, gives rise to the question of whether the comprehensive doctrine to be found in *A Theory of Justice* taints the early Rawlsian understanding (as we have just been discussing it) of the

personhood that is the immediate source or ground of the principles of justice. A brief review of the relevant issues will reveal that what is 'comprehensive' about the views expressed in *A Theory of Justice* has nothing to do with Rawls's early, Kantian understanding of the personhood that is the immediate source of the principles of justice; that is, the 'comprehensive doctrine' in *A Theory of Justice* has nothing to do with Rawls's *derivation* of the principles of justice in that early work. Rather, what is comprehensive – and therefore in need of reform – in *A Theory of Justice* has to do with the way in which persons are said to be motivated in *applying* and *adhering to* the principles of justice, once these principles have been derived. In other words, the Rawlsian self-criticism (as expressed in *Political Liberalism*) pertains to Rawls's earlier account of the *application*, and not of the *derivation*, of the principles of justice.

First of all, Rawls makes clear that his account of the 'original position' (which defines the initial choice situation for the personhood that is the source of the principles of justice) remains unaltered from *A Theory of Justice* to *Political Liberalism*.[37] Furthermore, the principles of justice that are derived via the original position are the same in *Political Liberalism* as they are in *A Theory of Justice*: 'All these elements [i.e. the principles that define Rawls's egalitarian liberalism] are still in place, as they were in *Theory*; and so is the basis of the argument for them.'[38]

Secondly, Rawls explains that the shift from *A Theory of Justice* to *Political Liberalism* was compelled largely by the inadequacy of his earlier account of the problem of stability, i.e. the problem of persons' motivations in continuing to adhere to the principles of justice (principles that have already been derived).[39] Everything that is new in *Political Liberalism* stems from Rawls's attempt 'to resolve a serious problem internal to justice as fairness, namely from the fact that the account of stability in Part III of *Theory* is not consistent with the view as a whole'.[40] That is to say, what is 'comprehensive' and Kantian (in the undesirable sense) in *A Theory of Justice* has to do with Rawls's account of stability – i.e. how the principles of justice are to be *applied* and *followed* by real persons – and not with his account of the *derivation* of the principles of justice on the basis of self-determining personhood (personhood in the original position).

Thirdly – and perhaps most importantly – what is Kantian about Rawls's earlier understanding (in *A Theory of Justice*) of the personhood that is the immediate source of the principles of justice (i.e. personhood in the original position) does not involve a 'comprehensive' or a non-political doctrine in the sense that Rawls criticizes in *Political Liberalism*. In fact, what is Kantian about Rawls's earlier understanding (in *A Theory of Justice*) of such personhood fully supports the later Rawlsian attempt to articulate a more completely political (not metaphysical)

conception of justice. Indeed, Rawls's later emphasis on the primacy of what is *practical* and *political* in his conception of justice can actually be understood as a reaffirmation and an intensification of his earlier Kantian understanding of the personhood that is the immediate source of the principles of justice. This is because what is Kantian about Rawls's earlier understanding is that such personhood is essentially self-determining and *not* dependent on any antecedently given metaphysical or teleological order of things. Such personhood is to arrive at the principles of justice based solely on *its own* practical and political purposes in organizing itself into a cooperative venture for mutual advantage. Indeed, the Rawlsian claim concerning the 'primacy of the practical' (a claim that becomes more pronounced in his later work, as he seeks to purge his theory of all remnants of a 'comprehensive' theory) is a specifically *Kantian* idea.[41] In sum: Rawls shifts from *A Theory of Justice* to *Political Liberalism*, not because the principles of justice and their derivation from personhood in the original position are wrong; rather, he shifts because his account (in *A Theory of Justice*) of why *real* persons are motivated to continue adhering to the (already derived) principles of justice (i.e. his account of stability) is not tenable, given the fact of political pluralism (i.e. a pluralism of reasons and motives).[42]

We shall see later just why Rawls felt compelled to criticize his own earlier account of persons' adherence to the principles of justice (i.e. his earlier account of stability). But for now, the main point to be grasped is simply that the earlier 'comprehensive', Kantian elements that Rawls later criticizes do not pertain to the personhood that is the immediate source of the principles of justice (personhood in the original position). Accordingly, what I shall say here concerning the personhood that is the immediate source of the principles of justice is valid for both the 'earlier' and the 'later' Rawls. We now turn to a further analysis of this personhood and why, for Rawls, it must differ conceptually from our own personhood.

III The problem of bias and the need to separate personhood into two types

The personhood that for Rawls is the immediate source or ground of the principles of justice is very much like *our very own* personhood in a significant respect: it is personhood that deliberates about and chooses principles of justice for itself, based on no antecedently given order of things but simply on its own purposes in organizing itself into a cooperative venture for mutual advantage. *Our* personhood is just like *that* personhood insofar as we are concerned about articulating principles of justice for our own practical, political purposes.

Now if there is this fundamental similarity between *our own personhood* and *the self-determining personhood that is the immediate source or ground of the principles of justice*, then why must we conceive of these two types of personhood as being any different at all? In other words, why shouldn't our *own* deliberations about justice and social arrangements lead directly to the set of principles that we are seeking? Why, for Rawls, do the principles of justice need to be derived by way of a conceptual detour through a fictional or hypothetical choice situation (the 'original position') involving personhood that is conceptually different from our own? The point can be expressed even more poignantly: if the personhood that is the deliberative source of the principles of justice is supposed to be essentially *self*-determining, and if we ourselves are already *self*-determining in the relevant respect (after all, we are engaged in the non-epistemological, non-metaphysical, political project of articulating principles of justice *for ourselves*), then why does Rawls find it necessary to establish a conceptual *distance* or *difference* between our own personhood and the personhood that is the deliberative source of the principles of justice (principles that we are supposed to *follow* as our criterion for making normative judgements about social arrangements)?[43]

For Rawls, *our* personhood cannot be understood as conceptually identical to the personhood that is the immediate source of the principles of justice that will be normative *for us*, because of *the problem of bias*. As Rawls notes, 'persons are not indifferent as to how the greater benefits produced by their collaboration are distributed, for in order to pursue their ends they each prefer a larger to a lesser share'.[44] Now in addition to being self-interested in this way, we as individual persons are differently situated *vis-à-vis* others in our social world: we have different natural endowments, we occupy different social positions, and we adhere to different conceptions of what is 'good' for us. If we were left to ourselves to deliberate about the principles of justice, these differentiating characteristics would naturally bias us and prevent us from reaching agreement. For Rawls, then, it is necessary to construct the fiction of an appropriate choice situation (an 'original position') within which personhood is *not influenced* by the particularizing characteristics that make us different as individuals and that bias us in our thinking about justice:

> Thus it seems reasonable and generally acceptable that no one should be advantaged or disadvantaged by natural fortune or social circumstance in the choice of principles. It also seems widely agreed that it should be impossible to tailor principles to the circumstances of one's own case. We should insure further that particular inclinations and aspirations, and persons' conceptions of their good do not affect the principles adopted. The aim is to rule out those principles that it would be rational to propose

> for acceptance, however little the chance of success, only if one knew certain things that are irrelevant from the standpoint of justice. For example, if a man knew that he was wealthy, he might find it rational to advance the principle that various taxes for welfare measures be counted unjust; if he knew that he was poor, he would most likely propose the contrary principle. To represent the desired restrictions one imagines a situation in which everyone is deprived of this sort of information. One excludes the knowledge of those contingencies which sets men at odds and allows them to be guided by their prejudices.[45]

In the initial choice situation (the 'original position'), then, personhood is understood as being ignorant about those particularizing characteristics (our particular natural endowments, our particular social positions, our particular conceptions of what is 'good') that differentiate us as individuals and that bias us in our understandings of justice. In the original position, a 'veil of ignorance' serves to 'filter out'[46] all knowledge of these different aspects of individual personhood, i.e. aspects that are arbitrary from the point of view of justice itself.

Now Rawls argues that persons in the original position are conceptually different from us insofar as they are to deliberate about justice from behind a 'veil of ignorance', devoid of all knowledge of contingent characteristics that make individual persons differently situated and differently biased. But must persons in the original position really operate behind a 'veil of ignorance'? Wouldn't it be sufficient simply to imagine that persons in the original position have been *instructed* to *ignore or disregard* all such differentiating characteristics?

Rawls indicates that such instruction would be inadequate. Persons in the original position must be understood to be actually ignorant of their differentiating characteristics, and for two related reasons. First, if such persons were not thought to be devoid of all such information, then it would be impossible to imagine them arriving unanimously at a single, determinate conception of justice:

> The restrictions on particular information in the original position are, then, of fundamental importance. Without them we would not be able to work out any definite theory of justice at all. We would have to be content with a vague formula stating that justice is what would be agreed to without being able to say much, if anything, about the substance of the agreement itself. . . . The veil of ignorance makes possible a unanimous choice of the particular conception of justice. Without these limitations on knowledge the bargaining problem of the original position would be hopelessly complicated. Even if theoretically a solution were to exist, we would not, at present anyway, be able to determine it.[47]

Secondly, if persons in the original position were not thought to be actually ignorant of their different situations in life, it would be impossible for them to escape the influence of arbitrary contingencies in their deliberations:

> Now the reasons for the veil of ignorance go beyond mere simplicity. We want to define the original position so that we get the desired solution. If a knowledge of particulars is allowed, then the outcome is biased by arbitrary contingencies. As already observed, to each according to his threat advantage is not a principle of justice. If the original position is to yield agreements that are just, the parties must be fairly situated and treated equally as moral persons. The arbitrariness of the world must be corrected for by adjusting the circumstances of the initial contract situation.[48]

Now to say that persons in the original position must be understood as being actually ignorant of their distinguishing characteristics as individuals is to say that they must be understood as being essentially different from ourselves for the purposes of the normative analysis being undertaken. After all, if persons in the original position were allowed to retain the kind of knowledge that we possess as individuals and were simply instructed to ignore or disregard everything that is arbitrary or contingent from the point of view of justice, then the persons in the original position would *not* be conceptually different from ourselves for the purpose of the normative analysis. For we, too, can be instructed (indeed, we can instruct ourselves) to ignore or disregard our knowledge of ourselves as particularized, differently situated persons. But as Rawls has indicated, such *self-imposed* ignorance or forgetfulness is not sufficient for the purpose at hand. Accordingly, the personhood that is the immediate source of the principles of justice must be *actually ignorant* and *actually different* (conceptually speaking, of course) from our own personhood; by implication, *ignorance in the original position cannot be self-imposed*, but must be *imposed on it externally and antecedently*, by a veil of ignorance that is already in place *before* such personhood begins to deliberate about justice.

For Rawls, personhood in the original position (and behind the veil of ignorance) is the normative, deliberative personhood that is the source of the principles that will serve as the criterion *for us* in our judgements about the justice of social arrangements. To the extent that such personhood lacks the knowledge that we have, such personhood is *conceptually different from* and *other than* our own personhood. In fact, for Rawls, if that personhood were not conceptually different from our own personhood, then the normative analysis would be impossible, since personhood in possession of all the knowledge that we have would be biased and unable to reach agreement on the principles of justice.

Of course, this talk about two different types of personhood should not cause us to reify that *other* personhood (i.e. personhood in the original position) or to treat it as an actually existing entity outside of us. As Rawls continually emphasizes, the notion of persons deliberating from behind a veil of ignorance in the original position does not refer to anything real, but is merely a 'device of representation'[49]: it serves as a model for what we regard as 'acceptable restrictions on

reasons available to the parties for favoring one political conception of justice over another'.[50] But while the notion of persons in the original position does not refer to any existent reality outside of us, it does refer to an *ideal* or a *model* of personhood that is essentially *different* from our own personhood. Thus, to the extent that personhood in the original position is *conceptually different* from our own personhood, it follows that, for Rawls, the deliberative personhood that is the immediate source of the principles of justice is different from our own (biased) personhood. Even if we are to regard persons in the original position as our own selves, only deprived of certain types of information, the Rawlsian project still requires this *conceptual* difference between *our* personhood and *that* personhood.

For Rawls, then, the veil of ignorance 'filters out' certain types of knowledge and renders the personhood in the original position ignorant of certain things and therefore conceptually different from our own personhood. But then what types of knowledge are to be filtered out, and what types are to be left available to persons in the original position?

As we have seen, Rawls's purpose is to assert a *normatively significant difference* between our personhood and the personhood that is the source of the principles of justice (personhood in the original position). That is, personhood in the original position must be constituted *differently* than our own personhood, and – as different – it is to be the source of principles that *we* are supposed to *follow* as we assess various social arrangements. Now the Rawlsian purpose in constructing the original position would be undermined by either one of two options: (1) if the veil of ignorance filtered out *all* information; or (2) if the veil of ignorance filtered out *no* information. If the veil of ignorance filtered out all information whatsoever, then persons in the original position would be completely ignorant and therefore would be unable to provide any principles of justice for us. Conversely, if the veil of ignorance filtered out no information whatsoever, then persons in the original position would be no different from us and, therefore, could not provide us with principles that differed from the ones that we would come up with for ourselves.[51]

Rawls aims to address this dilemma by distinguishing between (1) those circumstances of our existence that are merely arbitrary and contingent from the point of view of justice, and (2) those circumstances of our existence that concern all reasonable persons. For the purpose of constructing the original position, knowledge pertaining to (1) is to be filtered out by the veil of ignorance, but knowledge pertaining to (2) is to be allowed in. Accordingly, persons in the original position are to have no knowledge about the contingent, particularizing features that distinguish them from one another as individuals. However, persons in the original position are to have knowledge about their

general characteristics as persons, characteristics that concern all reasonable persons regardless of their different individuating features. Thus persons in the original position are allowed to know:

1 that they, like all other reasonable persons, value certain primary goods, 'things that every rational man is presumed to want,' such as basic 'rights and liberties, powers and opportunities, income and wealth';[52]
2 that they, like all other reasonable persons, are concerned about furthering their own particular interests and their own particular conceptions of the good (whatever these may turn out to be);[53]
3 the circumstances of justice, the basic 'conditions under which human cooperation is both possible and necessary': e.g. geographical proximity, a rough parity of mental and physical endowments among individuals, moderate scarcity, etc.[54]

According to Rawls, a theory of justice should 'assume as little as possible' at its basis; therefore, the persons in the original position are presumed not to share any 'extensive ties of natural sentiment'.[55] Rather, they are understood to be 'mutually disinterested'; this presumption is meant to 'insure that the principles of justice do not depend upon strong assumptions.'[56] Finally, in spite of their mutual disinterest, the persons in the original position are given a single, uniform task: to reach a stable agreement on a criterion or standard for assessing the basic structure of a self-contained social system existing in the circumstances of justice.[57] The criterion (or set of principles) that they arrive at will be binding *for us*, and will regulate our subsequent assessments of social arrangements.

Now in the original position as thus constituted, Rawls argues that personhood would arrive at two (lexically ordered) principles of justice, which he labels (1) the principle of liberty, and (2) the difference principle. These principles state, respectively:

1 'Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all'[58]; and
2 'Social and economic inequalities are to be arranged so that they are to the greatest benefit of the least advantaged'.[59]

The articulation of these two (lexically ordered) principles as the criterion for our normative judgements about social arrangements marks the culmination of Rawls's criteriological project as such. As Rawls summarizes:

> . . . the essential point is that despite the individualistic features of justice as fairness, the two principles of justice are not contingent upon existing desires or present social conditions. Thus we are able to derive a conception of a

> just basic structure, and an ideal of the person compatible with it, that can serve as a standard for appraising institutions and for guiding the overall direction of social change. In order to find an Archimedean point it is not necessary to appeal to a priori or perfectionist principles. By assuming certain general desires, such as the desire for primary social goods, and by taking as a basis the agreements that would be made in a suitably defined initial situation, we can achieve the requisite independence from existing circumstances. The original position is so characterized that unanimity is possible; the deliberations of any one person are typical of all.[60]

IV Problems with the Rawlsian account of the derivation of the principles of justice

Some of the most famous criticisms of Rawls have focused on questions concerning just how much[61] or how little[62] information Rawls presupposes in his derivation of the principles of justice. The criticism that I will offer will be somewhat different: I shall argue that the fundamental problem with Rawls's project does not turn on the question of how much or how little he seeks to filter out of the original position. The fundamental problem has to do with the fact that Rawls constructs an original position at all, and thereby separates (1) the personhood that is the immediate source of the principles of justice (personhood in the original position) and (2) our own personhood (the personhood that is supposed to be bound by the principles arrived at by personhood in the original position). That is to say, the fundamental problem does not have to do with the specific *content*, but rather with the very *form*, of the criteriological, contractarian strategy that Rawls attempts. In what follows, I shall try to show that the separation of personhood into two types leads to internal difficulties in Rawls's account of both the *derivation* and the *application* of the principles of justice. I shall begin with the former.

For Rawls, we need to construct an original position whereby persons constituted differently than ourselves are imagined to arrive at principles of justice that will be binding on us. We need to construct an original position with persons behind a veil of ignorance, because all of the specific knowledge that is available to us – if not filtered out – would bias persons' deliberations about the principles of justice. Such influences must be filtered out because they are arbitrary from the point of view of justice itself. Now a first line of criticism suggests itself rather easily: the Rawlsian approach presupposes the very thing that it aims to achieve. We are constructing an original position in order to arrive at principles of justice that will serve as a criterion for us and guide our normative judgements about social arrangements. But in our construction of the original position, we must decide which kinds of information

should and should not be filtered out; and therefore we must already know what kinds of information are arbitrary and not arbitrary from the point of view of justice. In constructing the original position for the sake of *arriving at* a criterion for our normative assessments about justice, we must already have available to us the kind of criterion we are seeking.[63] Thus the Rawlsian search for a criterion presupposes the very thing that it seeks to achieve.

Now the Rawlsian rejoinder is not too difficult to anticipate. Rawls would say that his own account freely admits – and, indeed, affirmatively presupposes – that we ourselves, in constructing an original position, already have available to us a valid and reliable 'sense of justice' that guides our conceptual construction of the original position. But if we already do have a reliable sense of justice, then it should not be necessary to construct an original position populated by persons that are essentially *different* from ourselves. We would already know how to make valid judgements concerning justice, and we would not need to appeal to any putative ideal or source that is *different* from ourselves. But Rawls has already argued that it *is* necessary to construct an original position, and that personhood in the original position *must be different* from our own personhood. More specifically, personhood in the original position must be ignorant of many things that we know about, because the kinds of thing that we know about bias us and prevent us from arriving unanimously at a reliable set of principles of justice. The problem here is that the bias that putatively taints our own deliberations about justice and putatively makes the construction of an original position *necessary*, also taints our very construction of the original position, and therefore makes an unbiased construction *impossible*.

The basic point is worth emphasizing. If a valid and reliable sense of justice is *already* available to us, then it is not necessary to construct an original position and thus not necessary to think of the personhood that is the source of the principles of justice as being conceptually *different* from our own personhood. On the other hand, if a valid and reliable sense of justice is not already available to us, then it is impossible to construct the original position that we need, for such construction requires us to know – in advance of having an actual criterion or standard available to us – what kinds of information are and are not arbitrary from the point of view of justice. The problem that makes it *necessary* to construct an original position (i.e. the problem of *bias*) is also a problem that makes it *impossible* to construct the kind of original position that Rawls requires. Thus the standard that Rawls is seeking is either not necessary (i.e. it is superfluous, since we are sufficiently unbiased) or it is impossible (since we are too biased).

By arguing that the personhood that is the immediate source of the principles of justice must be conceived as being different from our own

personhood, Rawls creates an awkward position for his criteriological project: as long as we *do not* already have the sought-for criterion available to us, we cannot arrive at it; and if we *do* already have the sought-for criterion available to us, then we do not need it. But the Rawlsian separation of personhood into two types is vulnerable to a second line of criticism as well. As we saw above, Rawls apparently was able to steer clear of epistemological and metaphysical issues because of his assertion of the primacy of the practical: the personhood that is the immediate source of the principles of justice is *self-determining* in the relevant respects. Because the goal of such personhood is simply to articulate principles for the purpose of *its own self-organization* into a cooperative venture for mutual advantage, there is no need to rely on any antecedently given epistemological or metaphysical order of things.

Now, contrary to Rawls's own self-understanding, I would like to argue that his very act of separating personhood into two types – our personhood and the personhood that is the immediate source of the principles of justice – implicitly contradicts his assertion of the primacy of the practical, and ensures that the personhood that is the (allegedly self-determining) source of the principles of justice is *not* self-determining in the sense that Rawls requires. With the Rawlsian separation, personhood that is the source of the principles of justice (personhood in the original position) is *not* self-determining in the relevant respects, because its putative interests, tasks, goals, self-understanding, and common nature are *given* to it by a source that is external to it, and given to it in accordance with a prior order of things that antecedently sets the terms of its deliberations (from 'behind its back', as it were). This is the antecedent order of things that *we ourselves* set up in advance as we construct the original position and decide what types of information should or should not be filtered out of it.

Now, on the face of it, this problem might not seem to be fatal, since the personhood that is the source of the principles of justice (personhood in the original position) was never meant to be an 'actual' personhood at all, and thus was never literally 'independent' and 'self-determining' in the full sense of the word. This is certainly a valid claim as far as it goes; but it is also true that, on Rawls's account, there is supposed to be a *real conceptual distinction* between our own personhood and the personhood that is the source of the principles of justice (personhood in the original position). Admittedly, personhood that is the source of the principles of justice is *not independent* and distinct from us in one sense (it is a construct of our own making); but for the purpose of the normative analysis, such personhood must be thought of as conceptually distinct from our own personhood, and therefore must be thought of as *independent* of our own particular desires, interests and

biases. According to my argument, it is this simultaneous independence and non-independence of the Rawlsian construct that has allowed Rawls to equivocate when talking about such personhood; and it is this equivocation that has made his derivation of the principles of justice only *seemingly* plausible. But when the equivocation vanishes, so does the plausibility.

One way to try to escape my criticism here is by arguing that the personhood that is the source of the principles of justice is not actually independent and self-determining in any ontological sense, but only *conceived* by us to be so. But if this is the case, then such personhood is also not self-determining in the sense that Rawls really *needs* it to be in order to avoid having to grapple with the 'non-political' concerns of epistemology and metaphysics. If such personhood is not actually independent and self-determining, then it is not *self-determining in the relevant respects*. This is because *we*, in constructing the original position, 'create' such personhood with a pre-existing common nature. This common nature pre-determines what that personhood may or may not take into consideration when deliberating in the original position. Insofar as *we* construct such an original position for *our own* political purposes, *we* might be said to be self-determining in the relevant respects. However, insofar as personhood in the original position must be conceived as *finding itself* already constituted with certain interests, tasks, goals, self-understandings, and a common nature *given to it* in accordance with an antecedently given order of things (i.e. the order that *we* have established for it), then *it* cannot be said to be *self-determining in the relevant respects*.

The same basic point can be stated from another angle. According to the Rawlsian account, there is no need for an extrinsic, non-political (epistemological or metaphysical) justification of the derivation of the principles of justice, because the source or ground of such principles is self-determining in the relevant respects; that is, the source or ground of the principles of justice is nothing other than the personhood whose own rights and duties are to be distributed in accordance with such principles. Now we can see that this is not so: personhood that is the source or ground of the principles of justice (i.e. personhood in the original position) is conceptually *different from* and *external to* the personhood whose relative rights and duties are to be distributed in accordance with such principles; this latter personhood is ourselves.

This is the reason why commentators on Rawls have been able to argue – correctly, I believe – that the Rawlsian project does not actually escape the need to grapple with non-political issues such as epistemology and metaphysics.[64] Rawls cannot avoid entanglement in issues of epistemology and metaphysics because his own project requires him to decide – in advance – what types of knowledge are to be included in the

'common nature' of persons in the original position. Here I shall focus on the overtly epistemological issues.

For Rawls, the goal in constructing the original position is to filter out all information about those contingencies that set actual persons at odds and that are arbitrary from the point of view of justice itself. Significantly, Rawls indicates that the reason why competing claims by particular individuals are arbitrary from the point of view of justice itself is that such claims can be reasonably rejected by other individuals.[65] Therefore, the question of what is or is not arbitrary from the point of view of justice (a putatively political question) is, for Rawls, bound up with the question of what kinds of claim can and cannot be reasonably rejected by different persons (an epistemological question). Furthermore, the assertion that a particular claim is reasonably rejectable (and therefore ought to play no role in constituting the 'common nature' of persons in the original position) is tantamount to saying that one should exercise skepticism with regard to that claim.

Now a Rawlsian defender might argue that the decision to disallow a particular claim from constituting part of the 'common nature' of persons in the original position does not imply any kind of skepticism with regard to that claim. Keeping certain kinds of claims 'out' of the original position, one might argue, has nothing to do with skepticism and everything to do with treating persons as free and equal by not imposing our particular views on them (i.e. particular views that they might reasonably reject). Unfortunately, this strategy still does not save the Rawlsian project from having to make implicit claims about epistemology and skepticism. After all, someone might insist that his/her particular (and even 'comprehensive') view of the good *does* belong 'in' the original position; furthermore, he/she might argue that such insistence is perfectly compatible with respecting persons as free and equal. To defend such insistence, the person only needs to hold as well that his/her particular view of the good is one that nobody could reasonably reject. Thus the construction of the original position *does* require – contrary to Rawls's own assertions – an implicitly skeptical epistemology with regard to those claims that are to be 'filtered out': the claims to be filtered out are the ones that could be rejected by other persons who, like ourselves, are reasonable. If *other* reasonable persons could reject such claims, then we *ourselves* might be epistemically *mistaken* in adhering to them.[66]

The construction of the original position implicitly requires an answer to epistemological questions concerning which kinds of claims we should and should not be skeptical about. It should also be clear by now why Rawls's requirement of an original position and his separation of personhood into two types do involve epistemological and/or metaphysical entanglements. In requiring an original position, Rawls makes

it necessary to decide *in advance* how persons' common nature (in the original position) would have to be constituted epistemically so as to rule out *all* disagreement and thus to *guarantee unanimity*. Thus the construction of the original position requires an epistemological rule that is general in application and that purports to decide questions of reasonable rejectability *in advance* of all the particular factual issues that might arise. By contrast, an approach to the question of justice that did not require an original position and that did not separate personhood into two types would not have to decide questions of reasonable rejectability *in advance*, and therefore would not have to include any general, epistemological 'rules of the game' for persons in the original position.[67]

Entanglements in epistemological and/or metaphysical questions could be avoided (1) if the personhood that is the immediate source of the principles of justice (personhood in the original position) really were self-determining in the relevant respects, or (2) if we ourselves (who are self-determining in the relevant respects, since we are searching for principles *for our own purpose*) were the immediate source of the principles of justice. But Rawls's criteriological, contractarian project – by requiring the separation of personhood into two types – guarantees that neither of these conditions is met: (1) personhood that can be said to be self-determining in the relevant respects (our own personhood) is not the immediate source of the principles of justice, and (2) personhood that is the immediate source of the principles of justice is not self-determining in the relevant respects. The Rawlsian account seems to achieve its stated goal, only because Rawls equivocates and blurs distinctions when discussing these two types of 'personhood'; but – according to Rawls's own contractarian, criteriological project – these two types of personhood need to be kept conceptually distinct.

V Problems with the Rawlsian account of the application of the principles of justice

Rawls's separation of personhood into two types not only creates internal difficulties for his account of the *derivation* of the principles of justice; it also creates problems for his account of how the principles of justice are to be *applied and adhered to* by persons in the real world (i.e. by us, who live and think on the *outside* of the original position).

As we have already seen, Rawls holds that a theory of justice should 'assume as little as possible' at its basis; accordingly, persons in the original position are presumed to be 'mutually disinterested' when deliberating about the principles of justice.[68] By the same token, Rawls presumes that persons in the real world who are to live in accordance

with the principles of justice (we ourselves) are also mutually disinterested.[69] Rawls begins with this presumption simply in order to ensure that the account of justice being proposed does not depend on any 'strong assumptions' or contentious claims about persons' values or moral motivations.[70] But, as I shall try to show, it is precisely this presumption of mutual disinterest – combined with Rawls's separation of personhood into two kinds – that requires Rawls to make strong assumptions and contentious claims when offering his account of stability (i.e. his account of why persons in the real world are motivated to continue adhering to the principles of justice).

Now let us assume, for the sake of argument, that the Rawlsian *derivation* of the principles of justice is problem-free. Even with this assumption in place, one can still reasonably ask: why should persons in the real world be motivated to continue *applying or adhering to* the principles of justice? If such persons are presumed to be mutually disinterested, then one cannot appeal to altruism or shared values in order to explain such motivation. The problem of motivation becomes particularly acute when one considers that persons in the real world might perceive the principles of justice to be in conflict with their own particular interests as individuals. How is it possible to explain persons' continuing adherence to the principles of justice, when there is the possibility of such conflict? This is the problem of stability.[71]

It is significant that persons in the original position can never be imagined to understand the problem of stability in the same way that persons in the real world understand it. This is because persons in the original position do not know what their 'particular interests' as individuals might be. Therefore, they cannot know what it is like to experience any conflict between the principles of justice and their own particular interests. Indeed, it is precisely because of this lack of knowledge and this lack of any perceived conflict that persons in the original position can be imagined to be capable of agreeing unanimously on the principles of justice in the first place.

The Rawlsian construction thus leads to the following results. By definition, persons in the original position cannot know what it is like to experience any conflict between the principles of justice and individual self-interest. And yet such persons are supposed to arrive at principles of justice that we (who *can* experience such a conflict) are supposed to follow. This creates obvious problems for Rawls. According to the internal requirements of the Rawlsian account, it is impossible for persons in the original position to be motivated in any way that might conflict with the principles of justice. By contrast, persons in the real world (we ourselves) might be motivated in any number of ways that conflict with the principles of justice. Because the first type of personhood specifies the ground rules to be followed by the second type of

personhood, there is a *motivational gap* that Rawls needs to fill. That is, Rawls needs to answer the following question: why would persons in the real world (persons whose particular interests might come into conflict with the principles of justice) be motivated to follow principles arrived at by persons who – by definition – do not and cannot know what it is like to experience any conflict between particular interests and the principles of justice? And why would persons in the real world continue to adhere to the principles of justice when their own particular interests *actually do* come in conflict with those principles?

According to Rawls's construction, persons in the original position – by definition – cannot provide an answer to this question, since persons in the original position cannot know *anything* of the conflict experienced by persons in the real world. As a result, Rawls must provide an answer to this question by appealing to some reason or ground taken from *outside* of the original position itself. Now Rawls's inevitable reliance on reasons derived from outside of the original position should make us suspicious; for as we have already seen, Rawls argues that the problem of self-interest and bias makes it necessary for us to distrust reasons or grounds that might be proffered by persons from outside of the original position. Thus once again we see that the internal dynamics of the Rawlsian project lead to serious difficulties: in order to explain why we should be motivated to continue adhering to the principles of justice, it is both necessary and impossible – on Rawls's own account – to appeal to reasons drawn from outside of the original position.

In order to escape this conceptual difficulty, one might argue that persons in the original position *can* give us good reasons for continuing to adhere to the principles of justice. One might argue that – although hypothetical persons in the original position know nothing of any *actual* conflict between particular interests and the principles of justice – they might be able to *remember* what it is like to experience such a conflict. Therefore, they can proffer good reasons for our continued adherence to the principles of justice. But this strategy will not do. For if persons in the original position possessed any *memory* of having *particular* interests, then that particularized memory would itself bias their thought in arriving at the principles of justice. Furthermore, it is not sufficient that persons in the original position be instructed to *ignore* their particularized memories for the sake of arriving at the principles of justice. After all, such deliberate ignoring would amount to a kind of *self-imposed* ignorance; and as Rawls has already argued, such self-imposed ignorance is inadequate for the task at hand. The ignorance – as well as the obliteration of all particularized memory – must be imposed *externally*.

Because of his presumption of mutual disinterest – coupled with his separation of personhood into two kinds – Rawls cannot answer the

question of stability (i.e. the question concerning our motivational grounds for adhering to the principles of justice) by reference to any notion of altruism or shared values. Accordingly, Rawls tries to argue that persons in the real world will be motivated to continue adhering to the principles of justice because such adherence is actually *desirable and good for us* as rational beings: 'acting justly is something that we want to do as free and equal beings'.[72] Furthermore, our desire to act justly is 'a desire to conduct oneself in a certain way above all else, a striving that contains within itself its own priority'.[73] Finally, persons' adherence to the principles of justice 'belongs to their good', and so 'the sense of justice aims at their well-being'.[74]

Now this attempt to fill the motivational gap between persons in the original position and persons in the real world rests on the following philosophical proposition: 'it is desirable and good for human beings to express their natures as free and rational beings, and thus to adhere to the principles of justice.' While Rawls does not rely on a notion of altruism to fill the motivational gap, he does rely on a philosophical doctrine of human nature that assumes more and is more contentious than his own stated intentions will allow. In other words, the motivational gap is filled only by a comprehensive doctrine – 'a Kantian interpretation'[75] – of what is good for human beings. And reliance on this kind of doctrine contradicts Rawls's stated *political* (as opposed to epistemological or metaphysical) project. It is precisely *this* Kantian aspect of *A Theory of Justice* that Rawls later criticizes and rejects as being part of a 'comprehensive' philosophical doctrine.

As indicated above (in Part II of this essay), Rawls's later work (e.g. as reflected in *Political Liberalism*) aims to overcome the inadequacy of his earlier account of stability and motivation. Thus in *Political Liberalism*, Rawls tries to show that real persons' continuing adherence to the principles of justice no longer needs to be explained in terms of what is 'good' for them as free and rational beings. Rather than relying on common motives, Rawls argues in *Political Liberalism* that persons in the real world may very well have different reasons and motives for respecting other persons in accordance with the principles of justice; but in spite of this pluralism, the basic Rawlsian conception of justice 'can gain the support of an overlapping consensus', where such a consensus is understood to consist of 'all the reasonable opposing religious, philosophical, and moral doctrines likely to persist over generations and to gain a sizeable body of adherents in a more or less just constitutional regime'.[76]

Now without delving too deeply into the Rawlsian account of the overlapping consensus, it is worth highlighting at least two difficulties with it. First of all, Rawls's account of what is and is not to be considered part of the 'overlapping consensus' of reasonable comprehensive

doctrines can and has been disputed by proponents of various 'reasonable comprehensive doctrines'.[77] In other words, Rawls's account of what should and should not be included in an overlapping consensus – like his account of what should and should not be included in the original position – seems to presuppose the very thing in question (namely, what shall count as a ground for reasonable agreement). Accordingly, it gives rise to the same kind of problems (epistemological and otherwise) that we have already highlighted with respect to Rawls's construction of the original position. Rawls has not really grappled with the problem of pluralism, but has only shifted it to another level.[78]

Secondly, even if Rawls's account of the specific *content* of an overlapping consensus is problem-free, the existence of such an overlapping consensus would itself be merely *contingent* and *arbitrary* from the point of view of the (Rawlsian) *reasons* given for the principles of justice. As we have already seen, Rawls argued that it was necessary to construct an original position, because persons in the real world could not be counted on to agree *unanimously* and *non-arbitrarily* on basic principles of justice.[79] Now the new Rawlsian account of an overlapping consensus seems to contradict this earlier argument regarding the original position, since the existence of an overlapping consensus is itself an arbitrary and contingent fact. Of course, it is possible to defend Rawls against this charge of self-contradiction by saying that – according to the new account – persons in the real world *can* be counted on to agree on basic principles of justice, even though they cannot be counted on to agree on *the reasons for* those basic principles. But if this is the new Rawlsian position (i.e. if Rawls now holds that we *can* count on general agreement about the principles of justice even as a contingent, empirical matter), then the construction of an original position – and the separation of personhood into two types – does not do any meaningful conceptual work any more. And yet – even in *Political Liberalism* – Rawls continues to hold onto the notion of an original position and the separation of personhood into two types.[80]

VI Criteriology and the implications of the radical primacy of the practical in Fichte

As we have seen, Rawls attempts to explain his principles of justice by reference to an imagined ideal choice situation wherein the personhood that is the source or ground of such principles is and must be conceptually different from our own personhood. I have tried to show that this basic dualism leads Rawls into internal difficulties on at least two different levels (i.e. with respect to both their derivation and their application). My immanent critique of Rawls has been informed by the

thought of Johann Gottlieb Fichte; however, in the spirit of immanent critique, I have refrained from relying on any claims other than those suggested by the internal dynamism of the Rawlsian project itself. Now, in this final section, I shall aim to show how the thought of Fichte supports and goes beyond the arguments that I have made with respect to Rawls.

From a Fichtean point of view, the Rawlsian failing can be expressed in a variety of ways. But perhaps the most straightforward way of expressing it in the present context is as follows: the fundamental problem with Rawls's criteriological, contractarian project resides in the Rawlsian attempt to imagine *in advance* what *particular positive content* (i.e. what particular claims or rules) free persons would agree upon for the purpose of organizing themselves into a self-sufficient social system. The problem with this approach is twofold:

1 any personhood that can be imagined by us in advance as being committed to any particular positive content is necessarily *not self-determining*; instead, it is always already constituted *by us* as being committed to some particular content that *we* have chosen for it;
2 any personhood that can be imagined by us in advance as being committed to any particular positive content is necessarily *not our own personhood*; this is because anything that subjectivity imagines or represents to itself is – strictly speaking – *other than* the free subjectivity that does the imagining or representing.

For Fichte, this twofold failure in the Rawlsian account has its roots in a more fundamental failure, namely the failure to recognize the *radical* primacy of the practical. For Fichte, this more fundamental failure can be overcome only if one also moves beyond all contractarian accounts that require the problematic separation of personhood into two types.

Unlike Rawls, Fichte does not seek to imagine *in advance* what *particular positive content* (i.e. what particular claims or rules) free persons would agree upon for the purpose of organizing themselves into a self-sufficient social system. Rather, the starting-point for Fichte is the exact opposite: insofar as personhood or selfhood is free and self-determining, it simply *cannot* be imagined in advance as being committed to any particular positive content whatsoever. Accordingly, it is not possible to specify in advance what kinds of claims one should or should not be skeptical about; in principle, personhood that is genuinely free and self-determining can be skeptical about any and all positive content whatsoever. Indeed, to be fully free for Fichte means to know that *no given content whatsoever* is necessarily determinative for one's thinking and/or acting. Conversely, to recognize that no given content is necessarily determinative for one's thinking and/or acting, is to recognize the radical priority of the practical.

Now while Fichte begins with a radicalized skepticism about all positive content as given, such skepticism does not lead to solipsism or nihilism, and does not rule out the possibility of genuinely *normative* conclusions; however, the normative conclusions to be drawn from Fichte's account do not depend on any claim concerning what kinds of positive content persons may or may not agree upon in advance. Rather, the normative force of Fichte's account is based on the proposition that – insofar as persons are radically free and self-determining – there is indeed *no* positive content at all that such persons can be imagined to agree upon in advance. Furthermore, once we grasp *what is required* for such radical, self-conscious freedom (i.e. once we grasp the hidden conditions of its possibility), we will see that the radically free self needs other free selves in order to be the self that it is, and that such selves must be understood as having *always already* agreed about something. Thus Fichte provides what might be considered 'a backwards social contract': persons must have always already agreed to something to the extent that they are self-conscious about their radical freedom and their ability to disagree about all positive content as given. In explaining this inversion, I also hope to shed some light on how Fichte overcomes the problematic Rawlsian dualism between our personhood and the personhood that is the source of the principles of justice (personhood in the original position), and how it amounts to a more complete enactment of the 'primacy of the practical', which Rawls correctly insisted on.

For Rawls, the problem of possible 'bias' or 'fallibility'[81] made it necessary to search for a criterion or standard whose source was essentially *other* than our own selfhood. As I suggested in Part IV, the problem that made the search for an external criterion necessary also made it impossible. The reason for the difficulty resides in the fact that Rawls must make two implicitly contradictory claims: on the one hand, the selfhood that is the source of the criterion being sought must be conceptually *different* from our own (potentially biased) selfhood; on the other hand, our very own (potentially biased) selfhood is *itself* ultimately responsible for properly identifying and applying this criterion.[82] From Fichte's point of view, the Rawlsian criteriological project represents a seemingly plausible application of the priority of the practical, but only because Rawls equivocates when discussing the two types of selfhood that are at issue: selfhood that provides the criterion (selfhood in the original position, or *noumenal* selfhood[83]) and the selfhood that is to be measured or tested in light of that criterion (our own, *phenomenal* selfhood). As part of his criteriological, contractarian project, Rawls needs to maintain this basic *dualism*, but also – as part of his insistence on the primacy of the practical – he needs to claim that our own practical, political selfhood is *self-measuring* or *self-testing* (and

thus he also needs to reject the dualism). For Fichte, a closer examination of the underlying issues will reveal that the very idea of self-measure or self-testing (i.e. the idea of applying any external criterion or standard to oneself) is internally problematic. We shall now consider these issues in more detail.[84]

First of all, it is important to note that genuine self-*measure* or self-*testing* is simply not possible for a self that is infallible – for there is no sense in measuring or testing that which is incapable of error. But if a self is fallible, what is the extent of such fallibility? For Fichte, there is in principle no reason why the self's fallibility does not extend to any attempted act of self-measurement or self-testing whatsoever. In other words, there is nothing to rule out the possibility that any alleged external or objective standard (or 'original position') to which the fallible (or biased) self might appeal in its act of self-measurement might be invalid or mistaken. Such a self, then, is not merely fallible, but *radically* fallible; its fallibility extends in principle to any attempt at identifying and/or applying a criterion for the purpose of self-measurement or self-testing.[85] We thus have what appears to be an insoluble problem: precisely because the self to be measured is fallible (for it would make no sense to measure or test an infallible self), there can be no guarantee that the standard by which the self seeks to measure itself is itself not mistaken or misapplied. This is the reason why the problem of bias, which for Rawls made it necessary to construct an original position, also makes it impossible to construct and interpret that original position without begging the very questions at issue.

For Fichte, the proper way to address the problem is to pay attention to our *awareness* of the problem *as* a problem; paradoxically, it is *our awareness* of the problem *as* a problem that contains the beginning of a genuine solution to it. But what exactly is contained in our *awareness* of the problem as a problem? For Fichte, to recognize that the self is radically fallible is to recognize that no given content or standard is *necessarily* determinative for the self's thinking and/or acting. In turn, to be aware that no given content or standard is necessarily determinative for the self's thinking and/or acting is to be aware that the self's thinking and/or acting is not determined by any external entity or force, but is radically free. The *meaning* of such freedom is susceptible to further elaboration; for now it need not mean anything more than that no given content (or idea or representation) necessarily imposes itself on us and forces us to accept it as our criterion or standard (i.e. we can be mistaken about any particular criterion or standard). For Fichte, the self's awareness of its *radical fallibility* thus coincides with the self's awareness of its *radical freedom*.

Now how are we to understand the self that is thus radically free and self-aware in its freedom? For Fichte, any account of the radically

free, self-aware self cannot be based on or derived from any given or determinate content. Precisely because the self is radically free, its being – and its being aware of itself as radically free – cannot be based on any given content or idea or representation. For Fichte, then, the self must be understood as nothing other than the activity of being aware of itself as radically fallible and free, undetermined by any given content. Two important qualifications are in order here. First, the awareness of oneself as radically fallible and free (an awareness that constitutes the self's very being) is necessarily a *non-representational* kind of awareness. After all, any representation belongs to that sphere of 'given' contents to which one may not appeal in defining the self as radically free. Secondly, the term 'awareness' must be used with caution here. For Fichte, the awareness that constitutes the self's being is nothing like any empirical awareness of a given, determinate content. Rather, it is an awareness that does not refer to or depend on any given content or fact (*Tatsache*) whatsoever. It is an awareness that is simply an activity (*Tathandlung*), namely the activity of being aware, in a non-representational way, of oneself as free and undetermined by any given content.[86]

Now what we have been describing thus far in our discussion of Fichtean selfhood is nothing other than what Fichte intends to convey through the first principle of his *Science of Knowledge*, namely the notion of the pure self, or *Ich = Ich*.[87] This activity of the self is alternatively described by Fichte as the activity of self-positing, or the activity of simple 'being *for* self'. The 'content' of the first principle of the *Science of Knowledge* is thus nothing other than the activity of self-positing, or being *for* oneself in a non-representational way. Here, the act of self-awareness and the content of the act fully coincide. All that the self is, is simply its own act of being for self, and all that is for the self, is simply its own selfhood as the act of being for self.[88] The self is not even a substance or thing that thinks (a *res cogitans*); it is nothing but the activity of thinking. It 'is an *act*, and absolutely [*absolut*] nothing more; we should not even call it an *active* something [*ein Thätiges*]'.[89] The Fichtean self is nothing other than the 'pure activity' of non-representational, non-substantialist self-awareness.

Here we have also hit upon the real meaning of the 'primacy of the practical'. Rawls was surely correct to insist on the primacy of the practical and to argue that our starting-point must be personhood or selfhood that is essentially free and self-determining. But Rawls implicitly violated his own initiative insofar as he conceived of such 'self-determining' selfhood as being outfitted *in advance* by a wealth of content (i.e. particular desires, goals and understandings) given to it by a source external to it (namely, by us). For Fichte, the real meaning of the primacy of the practical is that free selfhood is not and cannot be determined antecedently by *any* content that might be given to it, e.g.

desires, purposes, ideas, representations, or images. Thus Fichte finds it necessary to define the self as nothing other than the activity of being aware of oneself as free and thus as undetermined by any content whatsoever. However, Fichte's seemingly empty account of the self does not lead to solipsism or nihilism, but in fact entails a necessary relation to otherness, and this relation to otherness will have important normative implications.

As self-conscious of its radical fallibility and freedom, the self knows that no given content is necessarily determinative for itself, that no given content necessarily imposes itself on the self. However, one 'thing' that does 'impose' itself on the self is the fact that the self must always *come-to-be* aware of itself as radically fallible and free. The self's coming-to-be as a self-consciously fallible and free self always 'happens' to the self, apart from any deliberate or free choosing by the self. The self cannot deliberately and self-consciously choose its own coming-to-be-aware of itself as radically fallible and free (and thus cannot choose to come-to-be the self that it is), since – 'prior' to this coming-to-be – the self is 'not yet' a self-consciously free self at all. The self-consciously free self is what it is only to the extent that it emerges, or awakens, *out of* a 'prior' state of *not* being a self-consciously free self. Since the self *has not always been* the radically free and self-conscious self that it is, the self cannot be the totality of all that is, for coming-to-be necessarily implies some otherness. The self-positing self thus cannot be the totality of all that is, and there must be some *other* to the self, or a not-self (*Nicht-Ich*). This ineliminable otherness is what Fichte intends to convey through the second principle of his *Science of Knowledge*, the principle that the not-self is opposed to the self.[90]

The same point can be made in more Fichtean terms. To be a self is to be *for* oneself, and to be *for* oneself is to be given to oneself, and thus passive with respect to oneself. And one cannot be passive with respect to oneself (or in any respect at all), if the self were a pure, infinite, activity. That is to say, a pure, unconstrained, infinite activity – if it really were unconstrained and infinite – would never have the occasion to reflect *back* on itself or to be *for* itself, but would extend its activity without restriction or constraint into infinity – in which case it would be a blind, unreflected activity, and would not be an activity that is aware of itself, or for itself. Thus the very definition of the self as an activity that is purely for itself also implies some element of impurity, passivity and otherness. In order to be a self at all, the self needs an other in relation to which the self is the being-for-self that it is. Accordingly, the self-positing self cannot be the totality of all that is, and there must be some *other* to the self, or a not-self (*Nicht-Ich*).

But doesn't this account of the self and not-self lead to a contradiction? For Fichte, there is a contradiction, but it is an inescapable one,

since the very 'nature' of the self is to be in contradiction with itself.[91] After all, to be a self at all is to be always already *for* oneself or (what amounts to the same thing) to be self-relating, self-aware, self-positing, self-intuiting, self-measuring. But the condition of the possibility of the self's being the purely self-relating, self-positing self that it is, is that there be an other (not-self) *for* the self (i.e. an other 'within' the self's awareness). Thus the condition of the possibility of the self's being the purely self-relating, self-positing, self-measuring self that it is, is that it *not* be purely self-positing or self-measuring, but *in relation* to an other.[92]

In the earlier sections of this essay, I argued that Rawls equivocates or oscillates between two different kinds of personhood or selfhood: (1) *our own personhood* and (2) *personhood (other than ourselves) that is the source of the principles of justice*. Now in terms of Fichte's account, the first kind of personhood (our own biased, or *phenomenal*, personhood) is a kind of selfhood that is understood as being *not* purely self-related and self-positing, but other-related, conditioned, dependent, always already tainted by extraneous influences and contingencies; accordingly, this kind of selfhood seeks a standard or criterion whose source is *outside* of itself. By contrast, the second kind of personhood (personhood in the original position, or *noumenal* personhood) is a kind of selfhood that is understood as being purely self-relating and self-positing, uninfluenced by external contingencies, guided only by its own purposes; accordingly, this kind of selfhood is taken to be the non-arbitrary source of the sought-for criterion or standard. As we have seen, the Rawlsian project – on its own terms – required these two kinds of personhood to be conceptually distinct; and yet the Rawlsian project also had to deny the distinctness as well. For example, the biased, contingent, dependent selfhood (i.e. our own personhood) also had to be *not* biased, contingent, dependent, since it had to make valid normative judgements concerning how the original position was to be constructed; and conversely, the purely self-determining selfhood (i.e. personhood in the original position) also had to be *not* purely self-determining, since its interests and goals had to be set for it, behind its own back, in advance.

In my earlier remarks, I criticized Rawls for equivocating and oscillating between these two notions of personhood or selfhood. Now, in light of Fichte's account, we can see that *it is free, self-determining selfhood itself that does the oscillating and equivocating, and it does so necessarily*. The problem with Rawls, strictly speaking, was not that he failed to keep the two types of personhood conceptually distinct; the problem was that he thought it was possible to do so at all.

Now in addition to this argument about the internally contradictory, oscillating nature of the self, Fichte also develops an argument

about *intersubjectivity*. In the present context, the most important thing to note about Fichte's argument for intersubjectivity is that it does not involve any argument about what kinds of interests free persons (allegedly) all share, or what kinds of claims free persons (allegedly) can all agree upon. The argument is not grounded on any kind of givenness or positivity at all; rather, it is based on the opposite kind of claim: namely, that no given content whatsoever is necessarily determinative for the thinking and/or acting of a free self. As I shall try to show, Fichte's argument for intersubjectivity implies what might be called a 'backwards social contract'.

We have already seen that no particular content or claim is necessarily determinative for the self's thinking and/or acting; the self is radically free. We have also seen that – precisely because a free self is aware and must have come to be aware of its own radical freedom – there must be a not-self for the self. Now Fichte goes on to argue that this not-self must necessarily be *another free self*. One must keep in mind that the argument is presented as a *transcendental* argument in the Kantian sense. It is not about real, empirical selves or actual relations among persons or between persons and nature. It is an argument about the non-empirical conditions of the possibility of the self's ability to relate itself freely to any empirical objects whatsoever. For Fichte, the self simply could not be the self-consciously free self that it is, if there were no other free selfhood outside of itself. The crux of the argument, stated in the simplest terms, is as follows.[93]

While there must be a not-self for the self, this not-self cannot consist simply of 'nature'. Nature is that which sets no ends for itself, but rather has ends imposed on it externally, i.e. by free and purposive selfhood such as my own. Nature is simply the realm of the not-self insofar as it is given as an object to be controlled, consumed, dominated and transformed by me for my own purposes; that is, nature is that which is given to me from the outside, only to have its apparent independence canceled by me and integrated into my own purposive activity. Indeed, this very ability to cancel the apparent 'independence' of the natural object as given is a sign of my radical freedom. But to the extent that I merely consume nature or manipulate it to satisfy myself, I am also a slave to my passions or desires; for I consume and dominate nature as given in order to satisfy desires that, in turn, are merely *given to me* by my own (sensuous, internal) nature. Thus, to the extent that I am *only* a consumer in relation to what is other than me (i.e. in relation to the not-self), I am only a slave to my own naturally given desires, and thus not genuinely free (and thus also not capable of coming to be aware of myself as free).

In order to be genuinely free, and to be aware of my own freedom, I must not be a slave to my passions. I must not merely dominate what is other than me (the not-self) in order to satisfy my desires, but I must

let the other be; I must not simply impose my ends on it. But if the other that I 'let be' were itself just a piece of 'nature' and nothing else, and if I refrained from imposing my own ends on it, then the other would cancel my very existence *as a free being*. Such cancellation does not mean that the other would destroy me physically or biologically. But it does mean that the other would destroy my freedom, and thus would destroy me *as free*. After all, if the other were a merely natural, causally determined being, and if I refrained from asserting my own free purposiveness in relation to it, then my only relation to the other would be one of passivity; it would be a relation of *being-causally-effected* by it. Thus the other's existence in relation to me would cancel my existence *qua free being*. Therefore, if I am to be capable of refraining from imposing my own naturally given ends on the other (as I must be, in order not to be a slave to my passions), and if I am still to remain in existence *qua free being*, then the other must be capable of *preventing itself* from relating to me in a merely natural, causally determined manner. That is to say, the other must be capable of refraining from imposing *its* own naturally given ends on *me*, as *its* other; and this means that the other must be in a position to recognize that no naturally given content, as merely given, is necessarily determinative for its own acting. The other must be capable of canceling or overcoming its own merely natural existence, and this means that the other (i.e. the not-self) must be another free, purposive self.[94]

In his *Foundations of Natural Right*, Fichte explains the necessity of intersubjectivity by reference to the apparently impossible requirement of *finding* oneself as *free*. The problem is that the fundamental imperative that I have as a self is to *be* a self, or to be *for* myself, which means to *find myself as free*. But that means that my imperative is to find my own free efficacy as an *object*, and thus as finite, constrained, and determined – and that means as determined *by* something else (for all determinacy involves a relation to some otherness). But how can *free agency* find itself or see itself as thus *determined*? Of course, it cannot just find itself as determined by its *own* self, for it is precisely this *self-intuition* of the self that we are trying to explain – and to appeal to the self's seeing itself in the act of determining itself in order to explain how it sees itself as determined is to argue in a circle. But furthermore, the self cannot find itself or see itself as determined by a *mere object*; for in that case, the self would not be finding itself *as free*, and thus would not be finding *itself* at all. Fichte's claim is that the self can find itself as an object (as determined), only by finding itself as being-determined (summoned, called, or – in German – *aufgefordert*) to be self-determining by another self, and – more importantly – as *having already freely accepted* the call by the other self to be free and self-determining. And thus the self can find itself as free only by finding itself as having

always already agreed to, or accepted, a call or summons from another free self, even though – in a very real sense – the self was not deliberately and consciously present to itself or aware of itself in its acceptance of this call. After all, it is the self's acceptance of the summons or call (from the other free self) that serves to explain how the self comes to be aware of itself (or comes to find itself) as a deliberate, conscious, free self in the first place.[95]

Now without going further into the complexities of Fichte's derivation of intersubjectivity, it is possible to make some basic points about it and its relevance to the Rawlsian (or any similar) criteriological, contractarian project. Fichte's derivation of intersubjectivity implies that 'before' a self can be a self-conscious free self at all, it must always already stand in relation to another free self that allows it – and is allowed by it – to be free. Thus, even 'before' any free self can overtly begin reflecting on itself at all, it must have always already freely 'agreed' to stand in a relation of reciprocity or mutual recognition with another free self. 'Before' either self can be conscious of itself as free, both must have always already agreed to be free and to let the other be free. With this, it is possible to speak of something like a 'social contract' – or more accurately, a 'backwards social contract' – based simply on the self's radical freedom, i.e. the fact that truly free selves cannot be determined in advance to agree about any particular content or claim at all.

For Fichte, the problem for all normative social theory is not that we need to articulate specific grounds for reasonable agreement. The problem is that we have always already 'agreed' on relating to each other as free beings, but without having been conscious of such agreement and thus without knowing the 'terms' upon which such agreement was made. For how would one go about imagining the terms of such an agreement? Any attempt to imagine the specific 'terms' of this primordial 'agreement' between free beings (the 'agreement' that grounds their relating to each other as free beings) must presuppose that the persons are already aware of themselves as free beings. But for Fichte, their being aware of themselves as free beings (i.e. their finding themselves as free) is based on their having already (unselfconsciously) reciprocally 'agreed' to regard each other as free and thus as worthy of agreement. Thus any attempt to imagine the particular terms of this primordial agreement that lies at the basis of all intersubjectivity is futile; any such attempt must already presuppose *as having already occurred* that which one is trying to explain – namely persons' freely having come to be aware of themselves as the free beings that they are.

This account also indicates why the traditional social contractarian accounts (from Hobbes to Rawls) necessarily fail to explain what they seek to explain. All such accounts seek to imagine persons agreeing to the *basic terms under which they are to relate to each other as free*

persons per se. But insofar as such persons are imagined to be *bargaining* or *contracting* with each other at all, the imagined construct necessarily presupposes that the persons have *always already 'agreed' to something*, i.e. they have always already agreed to regard each other as free beings, capable of entering into agreements. While it is possible to imagine the terms under which persons might agree to relate to one another with regard to this or that particular matter, it is impossible to imagine the terms under which persons might agree to relate to (or recognize) each other as *free beings per se*. For the very idea of imagined contracting or bargaining presupposes that the persons have always already agreed to treat each other as *free beings per se*. Thus traditional contractarian approaches necessarily come on the scene too late to explain what they seek to explain.

Incidentally, this is also the reason why Rawls's account of the relation between his own political philosophy and the political philosophy of German Idealism misses the mark. Rawls argues that the project in *A Theory of Justice* and *Political Liberalism* is sufficiently sensitive to intersubjectivity and the social character of human beings, and therefore immune to the post-Kantian idealist critique of traditional contractarianism.[96] But according to my account, Rawls has misunderstood the real force of the German Idealistic argument. For Fichte and Hegel alike, a genuine sensitivity to intersubjectivity is not evidenced simply by the fact that one can imagine persons as having deep sentiments and commitments that bind them together.[97] Rather, a real sensitivity to intersubjectivity turns on what one *cannot* imagine. For Fichte and Hegel, the free self must understand itself as being *indebted* to other free selves *for its very awareness of itself as free*; accordingly, it is simply impossible to imagine free selves as agreeing to any particular or determinate terms under which they are to treat each other as free selves *per se*. To the extent that they are aware of themselves as free at all, they have always already agreed (unselfconsciously) to the very thing that one is trying to imagine. According to my argument, it is the imaginative, contractarian approach itself (and not just its Rawlsian application) that betrays an essential blindness to the primordial role of intersubjectivity at the basis of all self-conscious selfhood.

Now in spite of Fichte's emphasis on radically free, self-conscious selfhood, and on intersubjectivity as the hidden condition of its possibility, it does not follow that the Fichtean account is unable to provide any *determinate* direction for our normative and critical judgements about existing social and legal institutions. One must recall, of course, that the normative, critical force of Fichtean social theory cannot be based on any imagined story about the terms under which persons might have agreed to regard each other as free beings. For as we have seen, Fichte's starting-point is the proposition that free persons – insofar as

they are conscious of their freedom at all – must have always already reached 'agreement' with other free beings. Such (non-empirical, non-imaginable) agreement manifests itself for Fichte only *indirectly* in already existing social and legal institutions, such as property, contracts and criminal law.[98] And it is possible to criticize these institutions, but not because they allegedly fail to live up to the 'standard' of an imagined, hypothetical social contract. Rather, it is possible to criticize these institutions to the extent that the human self-understanding that they presuppose and foster as institutions fails to accord with the 'true' account of persons as radically free and self-determining, and as *intrinsically* related to other free beings. For example, property regimes can be criticized to the extent that they are based on and foster the notion that private property exists primarily for the purpose of satisfying our more-or-less animalistic 'natural' desires, without also – and more importantly – mediating our mutual recognition of each other as radically free beings.

In fact, based on the Fichtean premises of free selfhood and intersubjectivity, an argument can be made for something quite similar to Rawls's 'difference' principle. Recall that, for Fichte, the equilibrium of intersubjective recognition cannot be brought about by nature or force, but must be understood as resulting from the free, uncoerced activity of the selves involved. Now since such intersubjective recognition cannot be forced, or based on fear or oppression, it would seem to follow that there cannot be gross material inequalities between individuals within a given society. After all, the existence of gross material inequalities makes it possible – and in some cases all-too-tempting – for citizens to understand the existing social equilibrium as the result of nature or force or fear, rather than free, intersubjective recognition. And for Fichte's social theory (based as it is on the premise of radically free selfhood), what is crucial is not just that the social equilibrium be unforced or uncoerced, but that it also *be freely recognized* by the parties involved as unforced or uncoerced. However, the existence of gross material inequalities within a particular society tends to undermine the citizens' ability to recognize the existing institutions (such as property) as the products of their own free, intersubjectively mediated selfhood. With this, then, we have the beginnings of a normative, critical theory of society, not based on any contentious claims regarding what all reasonable individuals would agree to in a properly constituted original position, but based rather on the radical primacy of the practical, that is, based on the seemingly empty premise that free beings – precisely because they are free – cannot be imagined in advance as all agreeing to any particular thing at all.

Department of Philosophy, Fordham University, Bronx, New York, USA

Notes

1 It is worth noting that my immanent critique takes account of Rawls's more recent statements and revisions, including those found in *Political Liberalism* (New York: Columbia University Press, 1993) and *Justice as Fairness: A Restatement* (Cambridge, MA: Belknap Press of Harvard University Press, 2001). The faults that I identify in Rawls's theory are different from – and, to my mind, more serious than – the faults that Rawls seeks to correct in *Political Liberalism*, which hereafter will be cited as *PL*, and *Justice as Fairness*.

2 Thus my immanent critique of Rawls's theory of justice is also timely, for the first complete English-language translation of Fichte's own 'theory of justice' appeared only last year. See Johann Gottlieb Fichte, *Foundations of Natural Right According to the Principles of the Wissenschaftslehre*, ed. Frederick Neuhouser, trans. Michael Baur (Cambridge: Cambridge University Press, 2000).

3 See Michael J. Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982).

4 Sandel makes this sort of argument in various articles, and in his book, *Democracy's Discontent: America in Search of a Public Philosophy* (Cambridge, MA: Belknap Press of Harvard University Press, 1996).

5 John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), p. 4. Hereafter this work will be cited as *TJ*.

6 *TJ*, ibid., p. 4.

7 ibid.

8 ibid.

9 ibid.

10 As Rawls points out, the correct term is really 'lexicographical'; but that term is too cumbersome to use (*TJ*, 42–3).

11 *TJ*, ibid., p. 43.

12 *TJ*, ibid., p. 34.

13 ibid.

14 *TJ*, ibid., p. 39.

15 *TJ*, ibid., p. 40.

16 *TJ*, ibid., p. 3.

17 *TJ*, ibid., p. 9.

18 *TJ*, ibid., p. 13; emphasis added.

19 *TJ*, ibid., p. 584; emphasis added.

20 *TJ*, ibid., p. 20.

21 ibid.

22 *TJ*, ibid., p. 21.

23 *TJ*, ibid., p. 7.

24 *TJ*, ibid., p. 4.

25 *TJ*, ibid., p. 7. Thomas W. Pogge has noted (correctly, I believe) that Rawls – in a 1978 article, 'The Basic Structure as Subject' – seems to understand the term 'basic structure' more narrowly than he understands it in *A Theory of Justice*. However, this difference does not affect the present account of Rawls's 'principles of justice' as applied to the 'basic structure of society'.

When dealing with these terms, Pogge notes, the crucial point to grasp is that Rawls is focusing 'on the fundamental "rules of the game" and not on what moves players are morally free or constrained to make within a particular game in progress'. See Thomas W. Pogge, *Realizing Rawls* (Ithaca, NY: Cornell University Press, 1989), p. 26.

26 See, for example, Gerald F. Gaus, *Justificatory Liberalism: An Essay on Epistemology and Political Theory* (Oxford: Oxford University Press, 1996), pp. 3–4: 'Given the actual disagreement in our Western societies over liberal ideals, it is manifest that justificatory liberalism cannot explicate "publicly acceptable" principles as those to which each and every member of our actual societies, in their actual positions, actually assent. If that is the test of public justification, justificatory liberalism is most unlikely to vindicate substantive liberal principles. Justificatory liberals require a *normative theory of justification* – a theory that allows them to claim that some set of principles is publicly justified, even given the fact that they are contested by some. And this, in turn, appears to call for a moral epistemology, in the sense of an account of the conditions for justified moral belief, or at least justified adherence to social principles. . . . Remarkably, the adherents of justificatory liberalism not only fail to offer such an epistemology, but insist that abstaining from presenting one is fundamental to their position. John Rawls, for instance, maintains that "reasonable justification" is a "practical" and not an "epistemological" problem.'

27 *TJ*, pp. 11–12.

28 Of course, different social contractarians have proposed rather different ideas concerning the extent to which such personhood or selfhood is truly *self*-determining. For example, both Hobbes and Locke are social contractarians since both would agree that the principles of justice are to be derived from personhood itself insofar as such personhood chooses (or would choose) principles for itself and for its own purposes; however, both Hobbes and Locke also regard such personhood as *not fully* self-determining, since both regard such personhood as also partly constituted (e.g. by nature or by God) even *before* it begins to deliberate about the principles of justice to be chosen. For both Hobbes and Locke, this *prior* constitution of personhood determines the way in which personhood chooses the principles of justice for its own self-organization. Accordingly, the Hobbesian and Lockean accounts of the social contract are metaphysical (and contentious) to the extent that both rely on an account of 'nature' or 'God' that is allegedly given *prior to and independent of* the personhood that is supposed to be self-determining. Rawls, by contrast, seeks a non-epistemological, non-metaphysical, purely political foundation for his principles of justice, one that does not depend on any idea of a pre-given nature or God or teleological order of things. Accordingly, Rawls describes his contractarian approach as one that 'generalizes and carries to a higher level of abstraction' the traditional contractarian approaches of his predecessors (*TJ*, p. 11).

29 *TJ*, p. 11.

30 Michael Sandel suggests that this shift in Rawls's emphasis can be attributed to the 'communitarian' critique of Rawls's 'earlier' thought; Richard Rorty argues that this shift amounts to an abandonment of Rawls's earlier

'Kantian' account in favor of a 'Dewey-esque', pragmatist account. See Michael J. Sandel, 'Political Liberalism', *Harvard Law Review* (1994): 1765–94; and Richard Rorty, *The Virginia Statute for Religious Freedom* (Cambridge: Cambridge University Press, 1988), pp. 257–83. In my opinion, both Sandel and Rorty are probably wrong here. Contrary to Sandel, Rawls's heightened emphasis on the practical and political nature of his criteriological project can probably best be understood as a reworking of the problem of 'stability' in *A Theory of Justice*; contrary to Rorty, Rawls's heightened emphasis on the practical and political nature of his criteriological project can probably best be understood as an intensification – and not as an abandonment – of certain Kantian elements in *A Theory of Justice* (after all, it is a Kantian strategy to emphasize the primacy of the practical).

31 John Rawls, 'Kantian Constructivism in Moral Theory', *Journal of Philosophy* 77 (1980): 519. This article is reprinted in *John Rawls: Collected Papers*, ed. Samuel Freeman (Cambridge, MA: Harvard University Press, 1999), pp. 303–58.

32 ibid.

33 *PL*, pp. 29–35. See also John Rawls, 'Justice as Fairness: Political not Metaphysical', *Philosophy and Public Affairs* 14 (1985): 223–51 (reprinted in *John Rawls: Collected Papers*, pp. 388–414).

34 *PL*, ibid., p. 175.

35 ibid.

36 ibid.

37 *PL*, ibid., p. 7.

38 ibid.

39 *TJ*, p. 454.

40 *PL*, xv–xvi.

41 The *locus classicus* of Kant's claim concerning the 'primacy of the practical' is the 'Preface to the Second Edition' of the *Critique of Pure Reason*: 'Thus I had to deny *knowledge*, in order to make room for *faith*.' See Immanuel Kant, *Critique of Pure Reason*, ed. and trans. Paul Guyer and Allen W. Wood (Cambridge: Cambridge University Press, 1997), p. 117 (B edn p. xxx). The 'faith' to which Kant refers in this passage is a *practical* faith. Thus the point of the passage is to affirm the primacy of the practical: because we cannot have knowledge of God, freedom and the immortality of the soul as things-in-themselves, existing independently of our own activity in seeking to know them, we must affirm God, freedom and immortality as a *practical* matter, i.e. because they are necessary, as postulates, to our practical, moral activity.

42 My account of the shift from *A Theory of Justice* to *Political Liberalism* agrees in principle with the account given by Brian Barry. See Brian Barry, 'John Rawls and the Search for Stability', *Ethics* 105 (1995): 874–915.

43 This formulation of the question is meant to indicate – however proleptically – the internal difficulty that Rawls creates for himself in his criteriological, contractarian project. The basic problem is that the two types of personhood cannot be both self-determining (something that Rawls requires) and conceptually different (also something that Rawls requires).

Both our personhood *and* the personhood that is the immediate source of the principles of justice are supposed to be self-determining; however, Rawls's creation of a conceptual distance or difference between *our* personhood and *that* personhood ensures that each type of personhood is *not* self-determining: (1) *our* personhood is supposed to *follow* a criterion or standard that is given to it from a source that is essentially *other* than it, namely from *that* personhood; and (2) *that* personhood (in the original position) is *given* a narrow set of interests and concerns that it is allowed to consider in its deliberations about the principles of justice, and these interests and concerns are given to it from a source that is essentially *other* than it (i.e. they are given to it *by us*).

44 *TJ*, p. 4.
45 *TJ*, pp. 18–19.
46 For more on the role of a hypothetical social contract as a 'filtering' device, see Arthur Ripstein, 'Foundationalism in Political Theory', *Philosophy and Public Affairs* (1987): 115–37.
47 *TJ*, p. 140.
48 *TJ*, p. 141.
49 *PL*, 25ff. Rawls reaffirms this understanding of the original position in his most recent restatement. See Rawls, *Justice as Fairness*, pp. 14–18.
50 *PL*, p. 26.
51 Of course, the problem of filtering out an amount of information somewhere in between 'all' and 'nothing' is simply one way of expressing an age-old problem. Milton Fisk has recently described this problem with reference to the social contractarians' general dilemma: 'Which features of the human situation are to be left out in the attempt to isolate features of human nature? The first horn of the dilemma results from leaving out too much. Omitting all factors peculiar to a given epoch would imply that human nature involves only factors which can be shared by others, regardless of the social configurations they live through. Even if there were such transhistorical factors, it becomes a serious question whether they are sufficient to yield a set of principles on which a society can be built. The second horn of the dilemma results from including too much. Suppose the factors that express themselves in a set of social principles are internally related to a particular historical epoch. Then, treating these factors as if they defined a state of nature for humans means presenting historically specific factors as if they were independent of the accidents of history.' See Milton Fisk, 'History and Reason in Rawls' Moral Theory', in *Reading Rawls: Critical Studies on Rawls' 'A Theory of Justice'*, ed. Norman Daniels (Stanford, CA: Stanford University Press, 1989), p. 53.
52 *TJ*, p. 62.
53 *TJ*, p. 127.
54 *TJ*, pp. 126–7.
55 *TJ*, p. 129.
56 ibid.; see also *TJ*, pp. 144–5.
57 *TJ*, 7ff. and *TJ*, 126ff.
58 *TJ*, p. 302.
59 ibid.

60 *TJ*, p. 263.
61 For example, Robert Nozick accuses Rawls of assuming too much in his construction of the original position and his derivation of the principles of justice. See Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), pp. 183–231.
62 For example, Sandel criticizes Rawls for proposing an overly 'thin' theory of the good. See *Liberalism and the Limits of Justice*.
63 Robert Amico has made a similar point with regard to putative searches for epistemological criteria. In searching for an epistemological criterion, one sets up an impossible task: namely, the task of determining what will count as proper criteria of knowledge and what will count as proper instances of knowledge, yet without having any actual criterion of knowledge available in advance. See Robert Amico, *The Problem of the Criterion* (Lanham, MD: Rowman & Littlefield, 1993), pp. 112–15.
64 To give only two examples: Gerald Gaus has argued that Rawls cannot ultimately avoid presupposing some answers to epistemological questions, while Michael Sandel has argued that Rawls cannot ultimately avoid presupposing some answers to metaphysical questions. See Gaus, *Justificatory Liberalism*, and Sandel, *Liberalism and the Limits of Justice*.
65 *TJ*, pp. 136–42.
66 The argument that I have put forth here is similar to one proffered by Brian Barry. See Brian Barry, *Justice as Impartiality*, Volume II of *A Treatise on Social Justice* (Oxford: Oxford University Press, 1995), pp. 168–83.
67 If persons could deal with each contested issue freely as it arose and did not have to be bound by general epistemic rules laid out for them *in advance*, then they could be said to genuinely *self*-determining. But personhood in the original position is not this kind of personhood.
68 *TJ*, p. 129.
69 Of course, Rawls often insists that this presumption certainly does not rule out the possibility that persons in the real world might actually share very deep and lasting feelings of commitment and altruism. See *TJ*, p. 129.
70 *TJ*, p. 129; see also *TJ*, pp. 144–5.
71 Part III of *A Theory of Justice* is dedicated to this problem.
72 *TJ*, p. 572.
73 *TJ*, p. 574.
74 *TJ*, p. 476.
75 ibid.
76 *PL*, p. 15.
77 See, for example, Leif Wenar, '*Political Liberalism*: An Internal Critique', *Ethics* 106 (1995): 32–62.
78 ibid., p. 57.
79 *TJ*, pp. 140–1.
80 *PL*, p. 7.
81 In general, the term 'fallibility' may be more appropriate in an epistemological context, while 'the capacity for being biased' may be more appropriate in the context of political theory. In the present context, I shall use the notions of 'fallibility' and 'the capacity for being biased' interchangeably. Applied to the Rawlsian project, a fallible or potentially biased self is

simply one that might make normative judgements contradicting the principles of justice that would be agreed upon by (non-biased) persons in the original position.

82 Here I shift from using the word 'personhood' (the word that Rawls uses) to using the word 'selfhood' (the word that Fichte uses). This shift in terminology is theoretically insignificant, since the two words denote the same thing in the present context.

83 Not surprisingly, as part of his dualistic approach, Rawls also refers to selfhood in the original position as 'noumenal' selfhood. See *TJ*, p. 255.

84 My account here reiterates some of the basic points that I have made elsewhere in connection with Fichte. See my 'Self-measure and Self-moderation in Fichte's *Wissenschaftslehre*', in *New Studies in Fichte's 'Grundlage der gesamten Wissenschaftslehre'*, ed. Daniel Breazeale and Tom Rockmore (Amherst, NY: Humanity Books, 2001), pp. 81–102.

85 Fichte is thus willing to address the possibility of radical doubt; by contrast, Rawls thinks that it is necessary – and possible – to distinguish between claims that can and cannot be reasonably rejected (and he claims that one can do so without any reliance on epistemology). As I have suggested above, Rawls's opting for *moderate* skepticism made it necessary for him to propose (however implicitly) a positive epistemological account of what we should and should not doubt. Thomas Nagel seems to be aware of this necessity when he talks about the need to exercise 'epistemological restraint' when thinking about principles of social cooperation. See Thomas Nagel, 'Moral Conflict and Political Legitimacy', *Philosophy and Public Affairs* 16 (1987): 215–40. For Fichte, proper 'epistemological restraint' requires that one *refuse* to distinguish in advance between what can and cannot be reasonably rejected. For Fichte, the free being – by virtue of its radical freedom – is capable of questioning or rejecting any determinate claim or proposition whatsoever. For this reason, I (following Fichte) also believe that T. M. Scanlon's account of what one can or cannot reasonably reject is flawed. See T. M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Belknap Press of Harvard University Press, 1998).

86 Fichte's crucial distinction between a given content or fact (*Tatsache*) and an activity (*Tathandlung*) comes from the First Introduction to his *Grundlage der Wissenschaftslehre*. See J. G. Fichte, *The Science of Knowledge*, ed. and trans. Peter Heath and John Lachs (Cambridge: Cambridge University Press, 1970), p. 21. This distinction also helps to underscore the way in which skepticism about all types of givenness coincides with the radical 'primacy of the practical' in Fichte.

87 Fichte, *The Science of Knowledge*, p. 93.

88 As Fichte puts it, '*To posit oneself* and *to be* are, as applied to the self, perfectly identical. Thus the proposition, "I am, because I have posited myself" can also be stated as: "*I am absolutely* [*schlechthin*], *because I am.*" ' See Fichte, *The Science of Knowledge*, p. 99. Fichte's description of the self-positing self as 'absolute' [*absolut*], and the translation of the German '*schlechthin*' as 'absolute' or 'absolutely', can be misleading. A better translation would probably be 'simply'. Saying that the self 'absolutely' or 'simply' posits itself is not an attempt to infinitize the self, but

rather an attempt to express the radicalness of the self's fallibility. To say that the self 'simply' or 'absolutely' posits itself is to say that the self is so radically fallible as a knower that it is absolutely unable to explain itself or (what amounts to the same thing) explain its awareness of itself by appealing to any simple 'fact' or 'state of affairs' that it allegedly knows to exist 'objectively' and apart from its own selfhood.

89 Fichte, *The Science of Knowledge*, p. 21.

90 ibid., p. 102. See also J. G. Fichte, *Foundations of Transcendental Philosophy (Wissenschaftslehre) novo methodo (1796/99)*, ed. and trans. Daniel Breazeale (Ithaca, NY: Cornell University Press, 1992), pp. 121–33.

91 Hegel (following Fichte) expresses it this way: 'Here, we are not one-sidedly within ourselves, but willingly limit ourselves with reference to an other, even while knowing ourselves in this limitation as ourselves. In this determinacy, the human being should not feel determined; on the contrary, he attains his self-awareness only by regarding the other as other. Thus, freedom lies neither in indeterminacy nor determinacy, but is both at once.' G. W. F. Hegel, *Elements of the Philosophy of Right*, ed. Allen W. Wood, trans. H. B. Nisbet (Cambridge: Cambridge University Press, 1991), p. 42.

92 We have here, then, something like a 'transcendental' argument (in the Kantian sense) for the kinds of claims made by proponents of 'critical legal studies'. To give only one example, Duncan Kennedy argues that our legal understandings of persons oscillate, without the possibility of final resolution, between the extremes of persons-as-egoists and persons-as-altruists. On the basis of the account just put forth, one can offer a reason why such perpetual oscillation is necessary and unavoidable: by its very nature, the person or self is simultaneously both *self-related* and *related-to-an-other*. See Duncan Kennedy, 'Form and Substance in Private Law Adjudication', *Harvard Law Review* 89 (1976): 1685ff.

93 I speak of stating the argument 'in the simplest terms', but the argument is rather counter-intuitive, and thus not simple at all. However, I do believe that the argument that I put forward here – which echoes Hegel's language in the *Phenomenology of Spirit* – is one of the more intuitive ways of thinking about the Fichtean derivation of intersubjectivity. See G. W. F. Hegel, *Hegel's Phenomenology of Spirit*, trans. A. V. Miller (Oxford: Oxford University Press, 1977), pp. 109–10.

94 This argument, of course, implies a necessary *reciprocity* between the two selves. I cannot be a free self (i.e. I cannot overcome mere servitude to my passions) unless the other is also a free self (i.e. unless it is capable of canceling its own merely natural existence). And the converse is also the case: the other cannot be a free self (i.e. it cannot overcome its servitude to its passions and cancel its own merely natural existence) unless I, too, am a free self (i.e. unless I, too, cancel my merely natural existence and overcome mere servitude to my passions). Both Fichte and Hegel have emphasized the necessary reciprocity. For more on this topic, see Robert R. Williams, *Recognition: Fichte and Hegel on the Other* (Albany, NY: SUNY Press, 1992).

95 Fichte's argument appears in Section 3 (Second Theorem) of his *Foundations of Natural Right According to the Principles of the Wissenschaftslehre*, pp. 29–39.

96 *PL*, pp. 285–8. Similarly, sympathetic commentators on Rawls have stated or implied that Rawls escapes the German Idealist critique, since he is sufficiently sensitive to the social character of human existence. But these defenders, like Rawls himself, fail to address what I take to be the real point of the critique. See T. M. Scanlon, 'Rawls' Theory of Justice', in *Reading Rawls*, p. 177; J. B. Hoy, 'Hegel's Critique of Rawls', *Clio* 10 (1981): 409–11; and Chandran Kukathas and Philip Pettit, *Rawls: 'A Theory of Justice' and Its Critics* (Cambridge: Polity Press, 1990), pp. 122–3.

97 *TJ*, p. 129.

98 Hegel also takes these institutions as his starting-point. See Hegel, *Elements of the Philosophy of Right*, pp. 73–132.