

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

John Rawls: Between Two Enlightenments

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation	Frazer, Michael. 2007. John Rawls: Between two enlightenments. Political Theory 35, no. 6: 756-780.
Published Version	doi:10.1177/0090591707307325
Accessed	September 24, 2017 5:02:45 PM EDT
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:3342972
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)

The final, definitive version of this article was published in *Political Theory* 35:6, December 2007, pp. 756-780 by SAGE Publications, Inc. All rights reserved. © SAGE Publications, Inc. Available at http://online.sagepub.com.

John Rawls: Between Two Enlightenments

Michael L. Frazer, Harvard University

***Abstract**:* John Rawls shares the Enlightenment's commitment to finding moral and political principles which can be reflectively endorsed by all individuals autonomously. He usually presents reflective autonomy in Kantian, rationalist terms: autonomy is identified with the exercise of reason, and principles of justice must be constructed which are acceptable to all on the basis of reason alone. Yet David Hume, Adam Smith and many other Enlightenment thinkers rejected such rationalism, searching instead for principles which can be endorsed by all on the basis of all the faculties of the human psyche, emotion and imagination included. The influence of these sentimentalists on Rawls is clearest in his descriptive moral psychology, but I argue that it is also present in Rawls's understanding of the sources of normativity. Although this debt is obscured by Rawls's explicit "Kantianism," his theory would be strengthened by a greater understanding of its debts to the sentimentalist Enlightenment.

***Keywords**:* John Rawls, Immanuel Kant, David Hume, psychology, normativity

The Legacy of Two Enlightenments

Rawls as Heir of the Enlightenment(s)

John Rawls places himself firmly within the Enlightenment tradition by insisting on the right and responsibility of all individuals to reflect on the social structures which govern their lives. Rawls's goal is to formulate principles for the structuring of a society which can be reflectively endorsed by all its citizens. All human beings, he recognizes, share a capacity for

introspection, the ability to reflect upon their own thoughts and deeds in order to determine whether they ought to continue as before, comparing how things are actually done to standards of how they ought to be done. The specific form of moral reflection which Rawls investigates involves taking such a perspective on society's basic structure, and the relevant moral standards for this sphere are called principles of justice. Any element of society's basic structure, Rawls insists, is liable to rejection upon reflection if we conclude that it is unjust. Our standards of justice, like all our moral standards, are then themselves subject to revision upon reflection, and then further revision upon further reflection. Eventually, we may reach the conclusion that some of our standards are unlikely to be revised any further. We then treat these standards as authoritative. Rawls calls them our considered convictions in reflective equilibrium.[1]

When it is we who are the reflectors, it is we who determine our own moral and political standards. When we insist on reflective freedom—on the right and responsibility of all to reflect for themselves—we thus insist on the importance of autonomy, of self-legislation. The political metaphor of autonomy—so common that we often forget that it is a political metaphor—is a product of the eighteenth-century. The political revolutions of that time were grounded in a notion of literal, collective self-legislation through republican governance. The intellectual revolution of the same period, known as the Enlightenment, uses the enactment of legitimate positive laws by a self-governing people as a metaphor for the determination of principles of justice and morality through individual reflection.[2] Insofar as Rawls insists on the reflective autonomy of all individuals, he is continuing the revolutionary project of his Enlightenment forbearers.

Revolutionaries, however, always have trouble maintaining a united front. The study of eighteenth-century moral and political thought reveals that there were in fact many competing

Enlightenments, each with its own account of reflective autonomy. Although it is important not to oversimplify the intellectual diversity of the period, we can contrast two primary streams in the eighteenth-century analysis of moral and political reflection. The first, which I am calling *rationalist*, corresponds to our common conception of the eighteenth century as the "age of reason." The second, which I am calling *sentimentalist*, suggests an age, not of reason alone, but also of reflectively refined feeling. This is not to say that every moral and political thinker of the Enlightenment can be easily classified as exclusively rationalist or sentimentalist. Many of the greatest thinkers of the period—most notably Jean-Jacques Rousseau—evade such simple categorization. But there was clearly an ongoing debate in the eighteenth century over the nature of properly autonomous reflection—a debate in which many major thinkers took an identifiably rationalist position, and many others an identifiably sentimentalist one. David Hume and Adam Smith, for example, provide two different, but equally brilliant, defenses of Enlightenment sentimentalism, while Immanuel Kant provides perhaps the greatest single defense of Enlightenment rationalism. [3]

Although, in the eighteenth century, both rationalism and sentimentalism found many worthy advocates, the sentimentalist account of autonomous reflection is held in low esteem by most academic heirs of the Enlightenment today.[4] While a commitment to individual autonomy is still widely shared among liberal theorists, this commitment is most often understood in Kantian, rationalist terms: individual autonomy is identified with the individual exercise of reason, so principles of justice must be constructed which are acceptable to all on the basis of reason alone. The most prominent political philosopher of our time was not immune to this anti-sentimentalist attitude; when he wrote his masterwork, *A Theory of Justice,* Rawls explicitly presented his project as a Kantian one. The thesis of this essay, however, is that Rawls's work is

enriched by the fact that it stands between the rationalist and sentimentalist Enlightenments, drawing philosophical resources from both. Despite Rawls's own insistence to the contrary, his work owes as much to Hume and Smith as it does to Kant.

In order to clarify Rawls's debt to the sentimentalist as well as the rationalist Enlightenment, it is helpful to understand their competing theories of moral and political reflection as combining two separate elements. To use Hume's most famous distinction, they both offer a theory of what "is" and a theory of what "ought to be"—a descriptive moral psychology that explains what goes on when we engage in moral and political reflection and a theory of normativity which explains why the standards we reach through such reflection must be treated as authoritative. While sentimentalism describes reflection as a matter of feeling and imagination as well as cognition, rationalism describes reflection as a matter of rational cognition alone. While sentimentalism understands normativity as stemming from the reflective stability of a mind able to bear its own holistic survey, rationalism sees normativity as authoritative legislation by the faculty of reason—here identified with our true, free self. After these two areas of disagreement are further explicated in the remainder of this introduction, the essay will then proceed by examining Rawls's position on each of these areas of disagreement in turn. The second section of this essay will discuss the relationship between sentimentalism's description of our moral psychology and Rawls's description; the final section will then discuss the relationship between the sentimentalist theory of normativity and the theory (or, as we will see, theories) of normativity that Rawls provides. Although Rawls implicitly endorses much of sentimentalism's description of our moral psychology, he explicitly rejects the sentimentalist theory of normativity which might naturally accompany this descriptive psychology. Yet Rawls's own approach to

normative theorizing is far more compatible with the sentimentalist tradition than Rawls himself is willing to admit.

Two Theories of Reflective Autonomy

Although both the rationalist and sentimentalist Enlightenments are united in their endorsement of reflective autonomy,[5] they have different notions of what it means to legislate moral and political standards for oneself. They are divided on the nature of the self who is doing the legislating and the nature of the self who is obeying the standards so legislated. To use a Platonic locution, they disagree about which regime is proper within the individual soul. The rationalist theory of reflection separates the legislative faculties of the reflective mind—identified as "reason"—from the faculties that obey. The sentimentalist theory, on the other hand, sees the standards created in ethical reflection as products of the mind as a whole, and does not distinguish sovereign and subject aspects of the mind.

Admittedly, this reading of sentimentalism as a kind of democratic egalitarianism of the soul is in sharp contrast to the standard interpretation of Enlightenment sentimentalism. David Hume in particular is conventionally read as advocating a psychic regime as fully hierarchical as that of his rationalist opponents—disagreeing with them only as to which faculties are to be sovereign and which are to be subject. While rationalists from Plato onward maintained that reason is rightly the master and passion rightly the slave, Hume famously counters that "reason is, and ought only to be, the slave of the passions."[6] But this memorable bit of rhetoric distorts Hume's true view. Although philosophers may rightly distinguish the operations of reason from those of passion, Hume consistently maintains that the two are actually "uncompounded and inseparable."[7] It is true that Hume believes reason alone powerless to motivate action; it is in this

sense which reason is and ought to be passion's slave. Yet the sentiments which Hume describes as motivating moral action are not merely passions, but products of the mind as a whole, reason and imagination included. It is from passion alone that they get their motivational impetus, but moral sentiments are much more than mere impetus. So the contrast between rationalism and sentimentalism is best understood as the contrast between a hierarchical view of the moral soul on the one hand, and an egalitarian view on the other—an egalitarian view in which normatively authoritative standards are the product of an entire mind in harmony with itself.

Despite their hierarchical view of the proper psychic regime, Enlightenment-era rationalists considered their theory to be one of reflective autonomy because they identified themselves with the sovereign, legislative faculty and not with the subject faculties that obey its legislation. Although the other features of the mind and personality are plagued by contingency, reason deals only with necessary truths. Although my emotion, imagination and memory are all part of causal nexuses both natural and social, my reason is free. If I am to think of myself as free from natural and social contingency, I must think of my true self as purely rational. If my actions and my standards of action are to be truly my own, it is this real self which must be sovereign, legislating standards in reflection and dictating our behavior in practice.[8]

Unlike some of the more extreme rationalists of ancient times, Kant and his Enlightenment allies rarely denied that social and psychological contingencies are always responsible for much of our behavior. Rather than seek to extirpate the power of contingency from human life, they instead sought to bring all contingent forces under rational control, so that these contingent forces guide us to the very same standards and practices which reason necessarily and authoritatively demands.[9] Even if my norms or behavior are the product of social and psychological factors outside of authoritative reason, if these forces have been made to

comply with the dictates of my better, non-contingent self, then this behavior is rationally justified. So the Enlightenment rationalist position is generally Platonic, not Stoic; the passions are not to be banished from the psychic regime, but are to obey their superiors, and keep to their proper place. The duties of their station involve keeping quiet during the purely rational process of proper moral and political reflection, then obeying the rationally authoritative principles which emerge.[10]

Just as the passions take a subordinate place in the rationalist psychic regime, the study of these non-rational forces takes a subordinate place in rationalist moral and political theory. For rationalists, empirical anthropology is always subsidiary to the a priori metaphysics of morals. Only after reason has finished determining what standards we ought to follow can we then address the empirical question of how social and psychological contingencies may be better brought in line with reason's authoritative demands.[11]

Sentimentalism, by contrast, adopts a different attitude toward contingency, and identifies the true self with the whole self, contingent social and psychological elements included. Sentimentalist theorizing thus begins where rationalist theorizing ends—namely, with the empirical examination of what actually motivates us to follow our current standards and practices. Such motivations can be seen, the sentimentalists argue, to stem from moral sentiments—emotionally-charged products of our psychological makeup and social context as well as our rational cognition. The faculty of sympathy is central to their descriptive etiology of these moral sentiments. Sympathy is the bridge between the social and the psychological; it is the faculty by which inner mental states are shared among individuals. So the empirical social-psychology of reflection offered by sentimentalism can be understood largely in terms of the reflective expansion and correction of our sympathetic bonds to our fellow human beings.

Rawls's relationship to this richly descriptive social-psychological tradition of will be the subject of the next section of this essay.

Yet the sentimentalist account of reflection is not merely descriptive. The sentimentalists know that we not only approve and disapprove of our individual actions and our shared political practices, but also of our own sentiments of approval and disapproval. The fact that we can have higher-order moral sentiments—that we can approve or disapprove of our own approval and disapproval—allows for a process of reflection in which the mind as a whole repeatedly turns on itself as a whole, and winnows out those sentiments which cannot pass the test of reflection. Such psychologically holistic reflection leads us through a gradual progress of moral sentiments, as more and more of our contingently-given convictions are revised or rejected outright. Only those moral sentiments which endure when we reach reflective equilibrium can be treated as authoritative, for only minds in reflective equilibrium are capable, as Hume puts it, of "bearing their own survey."[12] Rawls's conflicted position on this theory of normativity as psychologically holistic reflective stability will be the subject of the final section of this essay.

Caring About Justice: Sentimentalism's Descriptive Moral Psychology

The Feminist Revival of Descriptive Sentimentalism

By the final decades of the twentieth century, academic moral and political thought was in the midst of a neo-Kantian moment. Anglo-American political philosophy was undergoing a revival under the leadership of John Rawls, while Jürgen Habermas was drawing European critical theory away from Marxism, and Lawrence Kohlberg was establishing moral psychology as a legitimate field for empirical, social-scientific study. All three, in their different ways, were admitted Kantians. The present work on Rawls grows out of a larger response to this neo-

Kantianism, a response which has achieved its greatest success so far in the empirical, social-scientific study of moral development.

Rawls's friend and Harvard colleague Lawrence Kohlberg famously sought to classify subjects asked to solve hypothetical moral dilemmas according to the degree of maturity shown in their ethical deliberations. He developed a classificatory scheme of six stages in which "each higher stage of reasoning is a more adequate way of resolving moral problems judged by moral-philosophic criteria."[13] While granting that his work relies on the validity of certain principles of philosophical ethics, "especially those of the formalist, Kantian tradition" (*PMD*, p. 279), Kohlberg nonetheless continues to maintain that no truly reflective ethical theorist could reject these essentially uncontroversial principles. Yet only those committed to a particularly Kantian form of Enlightenment rationalism would believe, as do Stage 6 subjects under Kohlberg's scheme, that "universal moral principles have a rational foundation" which establishes "that persons are ends in themselves and must be treated as such" (*PMD,* p. 176).

It can be an amusing exercise for moral and political philosophers to guess where on his scale of development Kohlberg would place their favorite canonical authors.[14] The Enlightenment sentimentalists, for example, seem to have many of the characteristics of Stage 3 subjects. Interestingly, Kohlberg does admit that Adam Smith offers "an excellent exposition of the Stage 3 elements of moral psychology," though he refrains from categorizing Smith as a Stage 3 subject per se.[15] "The sociomoral perspective of this stage," Kohlberg writes, "is that of an individual in relationships with other individuals. That person is aware of shared feelings, agreements, and expectations… An individual in this stage reasons by putting him/herself in the other person's shoes" (*PMD,* pp. 174-5).

In addition to describing Enlightenment sentimentalists, Kohlberg's portrait of the Stage 3 moral reasoner also describes, as Carol Gilligan observed, "the very traits that traditionally have defined the 'goodness' of women, their care for and sensitivity to the needs of others."[16] There has long been an association between sentimentalism and femininity. "The age-old split between thinking and feeling" Gilligan writes, "underlies many of the clichés and stereotypes concerning the difference between the sexes." Yet these stereotypes, Gilligan famously argues, point to "two modes of judging, two different constructions of the moral domain" (*DV*, p. 69)—a masculine mode based on abstract rules of justice or fairness, and a feminine mode based on care or sympathy for concrete individuals.

Feminist advocates of an "ethics of care" have frequently noted the affinity between their distinctive moral "voice" and that of the Enlightenment sentimentalists.[17] Annette Baier goes so far as to argue for a tradition of male moral philosophers who "should be given the status of honorary women," of whom Hume, despite his occasional pre-feminist moments of misogyny, is the greatest exemplar.[18] In keeping with Baier's analysis, Gilligan identifies the different voices of her study as "characterized not by gender but theme." Their association with gender, she admits, "is an empirical observation" (*DV*, p. 2)—an observation we now have very good empirical grounds for doubting.[19] Yet regardless of whether there is anything truly gendered about the rejection of moral rationalism, feminist ethicists are in large part responsible for launching the contemporary reclamation of Enlightenment sentimentalism, particularly in the field of empirical moral psychology.

Okin's Sentimentalist Interpretation of the Original Position

Given Gilligan's characterization of the rationalist moral voice as advocating an ethics of justice or fairness, it might be thought that Rawls's theory of "justice as fairness" would provide a paradigmatic example of such anti-sentimentalism.[20] Hume and Smith are indeed presented in a largely negative light throughout *A Theory of Justice*, albeit not insofar as they were sentimentalists, but only insofar as they are alleged to have been the intellectual ancestors of the utilitarian position against which Rawls frames his rival theory of justice.[21] The anti-sentimentalist interpretation of Rawls's descriptive moral psychology, however, has been convincingly refuted by Susan Moller Okin. "Whereas Rawls's theory is sometimes viewed as excessively rationalistic, individualistic, and abstracted from real human beings," she writes, "at its center (though frequently obscured by Rawls himself) is a voice of responsibility, care and concern for others."[22]

Okin admits that Rawls's Kantian intellectual heritage—with its "stress on autonomy and rationality as the defining characteristics of moral subjects," not to mention its "rigid separation of reason from feeling and refusal to allow feeling any place in the formulation of moral principles"—led Rawls to formulate his theory "in the language of rational choice" (*RF,* p. 231). Nonetheless, like all "the best theorizing about justice," Rawls's philosophy "has integral to it the notions of care and empathy, of thinking of the interests and well-being of others who may be very different from ourselves."[23] Caring about doing justice to our fellows, after all, involves caring about *them*, not merely insofar as they embody some abstract value of "dignity" or "humanity," but also insofar as they are concrete individuals and objects of our sympathetic sentiments. In opposition to many of her fellow feminists, Okin thus "questions the wisdom of distinguishing between an ethic of care and an ethic of justice" (*RF,* p. 247). Indeed, Enlightenment sentimentalists never oppose justice and care, or fairness and sympathy, as does

Gilligan. Like Okin, they instead maintain that our natural, emotional ties to each other play an integral role in our reflective commitment to justice.[24]

The hidden influence of the sentimentalist Enlightenment on Rawls's descriptive moral psychology is largely obscured by the contractarian thought experiment from which this theory is built, that of the original position.[25] Rawls asks us to imagine ourselves to be single-minded seekers of self-interest, creatures radically unlike the human beings that Hume and Smith describe, but not unlike the *homo economicus* of the rational-choice tradition.[26] Rawls describes the lack of "extensive ties of natural sentiment" on the part of the imagined actors of the original position as a special virtue of his thought experiment, one which allows him to draw on only premises which are "widely shared and yet weak. At the basis of the theory," he writes, "one tries to assume as little as possible" (*TJ,* pp. 111-112). To the weak assumption of mutual disinterest, however, Rawls adds a much stronger assumption: that in a fair deliberation over the principles of justice parties would be denied virtually all knowledge about themselves—from their race, sex and socioeconomic status to their religion and "conception of the good." Given that a "veil of ignorance" makes knowledge of one's particular identity impossible, in formulating principles of justice Rawls's imagined self-interested actors are forced "to take the good of others into account." A combination of mutual disinterest and selective ignorance can thus achieve "much the same purpose as benevolence" (*TJ,* pp. 128-129).

As Okin observes, moreover, this requirement of taking the good of others into account means that actors behind the veil of ignorance must differ from the *homo economicus* described by rational-choice theory in another important respect: such actors cannot rely on instrumental reason alone for purposes of the relevant deliberations, but also must possess the emotional and

imaginative capacity of well-developed empathy, something akin to what the English-language sentimentalists of the eighteenth century called "sympathy."[27] Okin writes:

> To think as a person in the original position is not to be a disembodied nobody. This, as critics have rightly pointed out, would be impossible. Rather, it is to think from the point of view of everybody, or every "concrete other" whom one might turn out to be… To do [so] requires, at the very least, both strong empathy and preparedness to listen carefully to the very different points of view of others (*RF,* pp. 245-248).

Consider, for example, Rawls's argument that parties in the original position would chose a principle guaranteeing a robust freedom of conscience for all. Facing the possibility that they will be deeply committed to a certain religious or philosophical vision of life, these actors must feel their way into the perspective of such committed believers, and consider the role that first principles play in the lives of those devoted to them. Upon doing so, Rawls argues, actors in the original position come to understand that any risk of not being allowed to live according to one's most cherished convictions cannot be compensated for by any degree of economic or political benefit (see, e.g., *TJ,* pp. 181-183). Such a rejection of the very sort of cost-benefit analysis characteristic of instrumental rationality cannot itself be the product of that very faculty, but must stem instead from an empathetic understanding of others, including those very different from oneself.

Such empathy is also relied upon when one considers the position of the economically worst-off when evaluating principles of distributive justice, as it is for virtually all subjects which parties in the original position are asked to consider. Rawls argues that this demand for empathy on the part of actors in the original position captures an important element in our intuitive understanding of the value of respect. "Mutual respect is shown," he writes, "in our willingness to see the situation of others from their point of view, from the perspective of their conception of their good… Thus to respect another as a moral person is to try to understand his

aims and interests from his standpoint" (*TJ,* p. 297). As Rawls explicates the term, "respect" here includes more than the mere regard for another's autonomy and rational interests which Kant would include under that idea, expanding to include what Elizabeth Anderson calls "consideration." Given the Kantian understanding of respect as a dispassionate regard for others' autonomy and rational interests, Anderson argues that "a different ethical concept—consideration—is needed to capture the engaged and sensitive regard we should have for people's emotional relationships."[28] Although Rawls's use of Kantian terminology to describe the intuitive values captured by the original position, the need for empathy on the part of agents in his thought experiment indicates that his attitude toward persons is as much a matter of sentimentalist "consideration" as it is one of Kantian "respect."

As for us, real-world individuals engaging in moral and political reflection outside the original position, we must combine all the empathetic capacities necessary for evaluating principles of justice from this imagined perspective with motivations sufficient to lead us to consider our society from such a perspective in the first place. "The motivation of the persons in the original position must not be confused with the motivation of persons in everyday life who accept the principles of justice and who have the corresponding sense of justice," Rawls writes. Should we, who are not behind the veil of ignorance, nonetheless choose to commit ourselves to the principles of justice which would be chosen by the imaginary, egoistic actors behind it, he realizes that our "desires and aims are surely not egoistic" (*TJ,* p. 128). For us, who know who we are and where our individual interests lie, to treat the perspective of the original position as authoritative—refraining from exploiting our obvious socioeconomic and political advantages—we need not only great empathy into the lives of those very different from ourselves, but also, as Okin writes, "a great commitment to benevolence; to *caring* about each and every other" (*RF,* p.

246). Under Okin's interpretation—and despite Rawls's own association of sympathy with the utilitarian position against which his own theory is framed—Rawls thus joins Hume and Smith in maintaining that "the capacity for enlarged sympathies" is "clearly required for the practice of justice" (*RF*, p. 237).

Rawls's Implicitly Sentimentalist Description of Moral Development

If the implicit sentimentalism of Rawls's descriptive moral psychology is obscured by the language of rational-choice theory used to describe the original position, this connection emerges more clearly in the later sections of *A Theory of Justice*. Here, as Rawls himself acknowledges, "aspects of the theory of justice are developed slowly from what looks like an unduly rationalistic conception that makes no provision for social values." Only after "the original position is first used to determine the content of justice" do we then see justice as "connected with our natural sociability" (*TJ,* p. 511). Nowhere is this more evident than in Chapter VIII in Part III of Rawls's book, which Okin laments has been "much-neglected" by those who see Rawls as a straightforward neo-Kantian (*RF*, p. 235). In this chapter, Rawls outlines a descriptive theory of moral development which bears certain similarities with Kohlberg's, but shows an emphasis on empathy and the moral sentiments which Okin sees as a turn away from Kant, and toward Adam Smith (*RF*, p. 237).

Characteristically, Rawls himself insists that his picture of moral development is derived from a tradition of "rationalist thought" which stretches from Rousseau and Kant to Piaget and Kohlberg. Rawls gives a very odd reading of this "rationalist" tradition, however, arguing that it views "the moral feelings as a natural outgrowth of a full appreciation of our social nature." He summarizes its position thus:

> We have a natural sympathy with other persons and an innate susceptibility to the pleasures of fellow feeling and self-mastery, and these provide the affective basis for the moral sentiments once we have a clear grasp of our relations to our associates from an appropriately general perspective… It is painful for us when our feelings are not in unison with those of our fellows; and this tendency to sociality provides in due course a firm basis for the moral sentiments…. (*TJ,* p. 403).

Although Rawls may follow Kohlberg in describing a movement from preconventional morality (what Rawls calls "the morality of authority") to conventional morality ("the morality of association") and to postconventional morality ("the morality of principles"), here moving from one stage to another involves an increased ability to see the world from the perspectives of others, combined with a growing sense of fellow-feeling which motivates one to make regular use of this ability (see *TJ,* pp. 410-411). Moral development builds from the care we first feel for friends and family; "wanting to be fair with our friends and wanting to give justice to those we care for is as much a part of these affections as the desire to be with them and feel sad at their loss." A sense of justice evolves as we extend this concern to all of society, perhaps all of humanity, and can be spurred by the development of emotional ties beyond our small sphere of intimates. "Thus in a well-ordered society where affective bonds are extensive both to persons and to social forms," Rawls writes, "there are strong grounds for preserving one's sense of justice" (*TJ,* pp. 499-500).

Even when Rawls describes "the morality of principles" which is the final stage in his scheme of moral development, the rationalist elements of his theory are combined with a strong sentimentalist strain. Although Rawls argues that a fully developed "sense of justice" must come to involve a commitment to acting justly for its own sake—and hence must "display an independence from the accidental circumstances of our world"—he also maintains that a commitment to justice for its own sake grows naturally out of concern for one's associates, and is continuous with our "natural sentiments" (*TJ,* p. 416).[29] Since Rawls sees the morality of

principle as a matter of reflectively endorsed moral sentiments inseparable from the experience of certain human emotions, his account of moral development actually owes at least as much to Hume and Smith as it does to Kant.[30]

Bearing our Own Survey: The Sentimentalist Theory of Normativity

Rawls on Hume's Theory of Normativity as Reflective Stability

No account of the development of our moral psychology could ever, by itself, justify our moral commitments; to believe otherwise is to confuse an empirical explanation of the origins of a value commitment with a demonstration of its genuine normative authority.[31] Yet once we accept a sufficiently sentimentalist description of our moral psychology—one which sees our moral commitments as reflective outgrowths of basic human emotions—a possible method for normatively justifying these commitments immediately suggests itself. Rawls recognizes this as "one main consequence" of his own implicitly sentimentalist account of moral development, for as soon as moral sentiments are closely linked to everyday emotions we see "that the moral feelings are a normal feature of human life" and that "we could not do away with them without at the same time eliminating certain natural attitudes." Drawing on his analysis of the sense of justice as an extension of the affective ties of friendship, Rawls thus writes that "persons who never acted in accordance with their duty of justice except as reasons of self-interest and expediency dictated… lack certain fundamental attitudes and moral feelings of a particularly elementary kind." Once we understand "what it would be like not to have a sense of justice—that it would be to lack part of our humanity too—we are led to accept our having this sentiment (*TJ,* pp. 428-429).

Rawls's idea of this reflective self-acceptance is directly parallel to the mode of normative justification which, in his lectures on moral philosophy, he attributes to Hume. Here, Rawls imagines a contemporary reader objecting to sentimentalist ethics as nothing more than descriptive moral psychology; Hume, under this view, "simply fails to address the fundamental philosophical question, the question of the correct normative content of right and justice." Yet to maintain such a position, Rawls counters, is "seriously to misunderstand Hume." Focusing on the conclusion of the *Treatise,* Rawls instead interprets Hume as maintaining "that his science of human nature… shows that our moral sense is *reflectively stable*: that is, that when we understand the basis of our moral sense—how it is connected with sympathy and the propensities of human nature, and the rest—we confirm it." [32]

Hume thus stands in direct opposition to Hobbes, Mandeville, Marx, Nietzsche, Freud and the other reductive debunkers of morality who often dominate discussions of descriptive moral psychology today. These debunkers have led many modern thinkers, Rawls included, to worry that we will come to "doubt the soundness of our moral attitudes when we reflect on their psychological origins" (*TJ,* p. 451). Christine Korsgaard, for example, expresses concern that certain psychological theories may adequately explain, from a third-person perspective, why individuals act morally while making it impossible for agents to justify such action from their own, first-person perspective unless "kept in the dark about the source of their own moral motivation."[33] If true, such a descriptive psychological theory would make the reflectively-informed normative justification of our moral commitments impossible, for, as Korsgaard writes, "to raise the normative question is to ask whether our more unreflective moral beliefs and motives can withstand the test of reflection." The fear of such theories is, she argues, why "we

seek a philosophical foundation for ethics in the first place: because we are afraid that the true explanation of why we have moral beliefs and motives might not be one that sustains them."[34]

If Hume's (or, for that matter, Rawls's) description of our moral psychology is correct, however, a complete understanding of the origins of our proper moral commitments can only help us "gain the peace and inward satisfaction of being able to bear our own survey."[35] Although such a defense of human morality cannot convince a committed Kantian, who is determined to find a morality binding on any rational being as such, the Humean can simply reply, with Rawls, that "beings with a different psychology either have never existed, or must soon have disappeared in the course of evolution" (*TJ,* p. 433).

Although Rawls is clearly aware of the possibility of such a normative vindication of his sense of justice, he nonetheless rejects it explicitly. "The fact that one who lacks a sense of justice… lacks certain fundamental attitudes and capacities," he writes, "is not to be taken as a reason for acting as justice dictates" (*TJ,* p. 428). Rather than being presented as a possible normative justification of his theory, the discussion of moral development in Chapter VIII of *A Theory of Justice* is meant merely to counter one possible objection to justice as fairness: namely, that it might prove unstable over time. "One conception of justice is more stable than another," Rawls writes, "if the sense of justice that it tends to generate is stronger and more likely to override disruptive inclinations." In order to achieve such stability, it is thus critical that "when institutions are just (as defined by this conception), those taking part in these arrangements acquire the corresponding sense of justice and desire to do their part in maintaining them… However attractive a conception of justice might be on other grounds," Rawls concludes, "it is seriously defective if the principles of moral psychology are such that it fails to engender in human beings the requisite desire to act upon it" (*TJ,* p. 398).

His description of moral development, however, makes it clear that Rawls's own conception of justice is not liable to this criticism; it shows "how justice as fairness generates its own support and… that it is likely to have greater stability than the traditional alternatives" (*TJ,* p. 399). This description of our moral psychology, however, is explicitly not intended as part of the reflective justification for Rawls's conception. "The main grounds for the principles of justice have already been presented," Rawls claims. "At this point we are simply checking whether the conception already adopted is a feasible one and not so unstable that some other choice might be better" (*TJ,* p. 441). This conception of empirical psychology's subsidiary place in moral and political theory is, as has already been discussed, characteristic of Enlightenment rationalism. Only after reason has finished the justificatory work of normative reflection do rationalists then examine how the rest of the human psyche can be made to conform to reason's authoritative demands.

Sentimentalist and Rationalist Sources of Normativity in A Theory of Justice

In the 1971 presentation of justice as fairness in *A Theory of Justice*, Rawls's own approach to the question of normative justification is highly ambiguous. On the one hand, Rawls puts forward a broad notion of reflective equilibrium fully compatible with sentimentalism; on the other hand, he also emphasizes a "Kantian interpretation" of the original position which suggests that the conclusions drawn from this thought experiment are justified in a rationalist manner.[36]

The idea of reflective equilibrium is introduced, not in the normative justification of justice as fairness per se, but rather in its descriptive formulation. Here, Rawls suggests that "one may think of moral theory at first… as the attempt to describe our moral capacity; or, in the

present case, one may regard a theory of justice as describing our sense of justice" (*TJ,* p. 41). By going back and forth between the commitments to which our already-developed sense of justice leads us and the philosophically precise conceptions of justice considered from the point of view of the original position, we can come to settle on a precisely articulated theory of justice that both matches our considered convictions and meets the test of philosophical scrutiny. Of course, a full process of reflective equilibrium does not merely describe "a person's sense of justice more or less as it is, although allowing for the smoothing out of certain irregularities;" the successful attainment of equilibrium might instead necessitate that one's initial sense of justice undergo a radical shift (*TJ,* p. 43). The normative authority of our ultimate moral commitments then derives from their reflective stability. "Justification rests upon the entire conception and how it fits in with and organizes our considered judgments in reflective equilibrium," Rawls writes. Philosophical justification is here "a matter of the mutual support of many considerations, of everything fitting together into one coherent view" (*TJ,* p. 507).

Such a theory of reflective equilibrium is wholly compatible with the sentimentalist tradition's understanding of moral and political reflection. Indeed, Rawls describes his theory in its initial stages as "a theory of the moral sentiments (to recall an eighteenth century title) setting out the principles governing our moral powers, or, more specifically, our sense of justice" (*TJ,* p. 44). If anything, the notion of reflective equilibrium in the sentimentalist tradition is even more robust than Rawls's own. Although Rawls sometimes seems to see the quest for reflective equilibrium as basically cognitive process involving the weighing of pre-philosophical beliefs against philosophical theories, the sentimentalist conception of reflection involves a holistic attempt to reach an equilibrium on which the faculties of reason, feeling and imagination can all settle in harmony.[37] Yet if Rawls's arguments successfully establish that his is the best available

theory of justice under cognitive reflective equilibrium, then it seems likely that his theory will capture our settled commitments under a broader, sentimentalist reflective equilibrium as well. Rawls's own implicitly sentimentalist descriptive moral psychology certainly indicates as such.

Yet Rawls also devotes a section of his book to "the Kantian interpretation" of justice as fairness (*TJ,* pp. 221-227), and then alludes to this interpretation throughout the rest of the work. Here, the combination of mutual disinterest and the veil of ignorance in Rawls's thought experiment are described, not as a functional substitute for our feelings of benevolence, but as capturing important features of autonomous rational choice. Parties behind the veil of ignorance are the analogues of what Kant would call our noumenal selves, choosing laws to govern their behavior without any reliance on natural, heteronomously determined contingencies. "The principles he acts upon," Rawls writes of such an autonomous agent, "are not adopted because of his social position or natural endowments, or in view of the particular kind of society in which he lives or the specific things that he happens to want" (*TJ,* p. 222). One important reason, Rawls argues, for essentially autonomous beings to commit themselves to the principles which they would choose behind the veil of ignorance is thus to free ourselves from the power of all natural and social contingencies. It is at this moment that Rawls departs most sharply from the sentimentalist movement on the subject of normative justification.

Yet any picture of Rawls as a Kantian on the subject is complicated by the fact that Rawls elsewhere presents "the connection between acting justly and natural attitudes" as one of three possible arguments for why the sense of justice is good for those who possess it. This (implicitly) sentimentalist line of reasoning is followed both by the (explicitly) "Aristotelian" argument that "participating in the life of a well-ordered society is a great good" and the (also explicitly) "Kantian" argument that "acting justly is something we want to do as free and equal

rational beings" (*TJ,* pp. 499-501). Although Rawls repeatedly emphasizes the debt his theory of justice owes to Kant, it is clear that "the Kantian interpretation" is merely one interpretation of the sources of its normative authority, and that other interpretations are possible, including sentimentalist interpretations.

The Two Enlightenments in Overlapping Consensus

In his later writings, rather than settling decisively on any of the interpretations of normative justification outlined in *A Theory of Justice*, Rawls simply decided to bracket the question of full philosophical and normative justification as such. The "political liberalism" of these later works "applies the principle of toleration to philosophy itself,"[38] thus opening the way for adherents of a variety of worldviews to participate in an overlapping consensus supporting justice as fairness. Rawls here argues that his theory of justice should be acceptable to many of those in a democratic society otherwise divided in their beliefs. The doctrines held by those participating in this overlapping consensus will likely be "comprehensive," meaning that they include "conceptions of what is of value in human life, ideals of personal virtue and the like," and hence, when we affirm them, "inform much of our conduct (in the limit of our life as a whole)."[39] Those participating in an overlapping consensus share political conceptions such as justice, but insofar as they affirm different fully and partially comprehensive doctrines, they will take opposing stands on other moral and philosophical matters.

Yet since political convictions "are also, of course, moral convictions" (*PL,* p. 119), an overlapping consensus necessitates substantive agreement on a number of important moral commitments. Rawls thus distinguishes an overlapping consensus from a *modus vivendi*; only in the former is political justice "affirmed as a moral conception" (*PL,* p. 168). Such shared moral

commitments, he argues, must be presented as a freestanding set of convictions, one which "formulates its values independent of non-political values and of any special relationship to them."[40] Political liberalism, with its conception of justice as fairness, is thus a moral module which "in different ways fits into and can be supported by various reasonable comprehensive doctrines that endure in the society regulated by it" (*PL*, p. 145). Since this module can mesh cleanly with many otherwise conflicting worldviews, the only individuals left outside a liberal society's overlapping consensus are those who "cannot support a reasonable balance of political values" (*PL*, p. 243).

It is easy to be led astray by Rawls's description of the balance of values in this moral module as "reasonable." An appeal to the reasonable certainly sounds like an appeal to a Kantian notion of morally authoritative reason. Rawls, however, sees the reasonable, not as a criterion of rationality which transcends opposing moral commitments, but as itself a moral commitment, and his theory explicitly "does not try to derive the reasonable from the rational" (*PL*, p. 52). The use of the term "reasonable" to designate acknowledgement of certain moral principles may seem unusual, but Rawls insists that "common sense views the reasonable but not, in general, the rational as a moral idea involving moral sensibility."[41] Rawls's political liberalism does not insist that "the reasonable is the whole of moral sensibility." It does assert, however, that the reasonable "includes the part [of morality] that connects with the idea of fair social cooperation" (*PL*, p. 51).[42] Each reasonable comprehensive doctrine in an overlapping consensus provides its own full normative justification for the authority of this part of morality.

In this way, it becomes "central to political liberalism that free and equal citizens affirm both a comprehensive doctrine and a political conception [of justice]" (*PL*, pp. 608-609). Once philosophy reaches its political limits, a theory of justice must still be vindicated by a

comprehensive doctrine. Rawls maintains that political philosophers must reconcile themselves to the inevitable pluralism of the modern nation-state, and hence help to build a regime of social cooperation including those with many diverse worldviews. Therefore, by necessity, modern political philosophy "proceeds from some consensus: from premises that we and others recognize as true, or as reasonable for the purpose of reaching a working agreement on the fundamentals of political justice."[43] Rawls makes a convincing case, Jean Hampton argues, that "whatever else political philosophy ought to involve," it should sometimes focus on how best to create a truly political liberalism, one which can serve as a shared commitment for those otherwise divided by moral, religious and philosophical differences.[44] Rawls's political liberalism is the sort of theory that we must turn to when, as Rawls himself puts it, "our shared political understandings… break down."

Rawls realizes, however, that we must also turn to political philosophy "when we are torn within ourselves" (*PL*, p. 44). If we cannot reflectively commit ourselves to a conception of justice which we can share with our fellow citizens, Rawls writes, we may "grow distant from our political society and retreat into our social world"; we may "feel left out… withdrawn and cynical."[45] It is for this reason that Thomas Nagel writes that the "ultimate aim of political theory" is to "justify a political system to everyone who is required to live in it."[46] Those with a well-defined, fully comprehensive worldview can often, at least hypothetically, be given such a justification through a skilled appeal to certain authoritative texts or traditions. Yet those seeking a full philosophical justification for liberalism must begin the work of justification from scratch. Specifically, Rawls maintains that they must construct what he calls an "Enlightenment liberalism." By this, he means a "comprehensive liberal and often secular doctrine founded on

reason," one capable of supporting the reasonable (in the moral, Rawlsian sense) through a direct appeal to the rational (*PL*, p. xl).

Even as he moves away from the comprehensive, Kantian justification for justice as fairness discussed in *A Theory of Justice*, Rawls thus shows himself to be deeply influenced by the sage of Königsberg. Kant's philosophy remains, for Rawls, the paradigmatic example of a comprehensive "Enlightenment" view which provides a full reflective justification for its conception of justice. Rawls here implies the rationalist Enlightenment was the only Enlightenment. A theory of justice, under Rawls's view, can thus either be political or metaphysical—built either from a consensus among existing worldviews or from necessary, rationally demonstrable truths.

Yet there is no reason to believe that the justification of liberalism's conception of justice requires appeal to a comprehensive system of categorical, *a priori* rules valid for any conceivable rational being as such—a metaphysical appeal, that is, of the sort that Kant and other Enlightenment rationalists thought necessary for the justification of ethical principles. This justification need only demonstrate that, all things considered, we have good reasons to commit ourselves to a liberal conception of justice, not that these reasons are discoverable *a priori* through the operations of pure reason. Just as important, the reasons in question need only be good for *us*—for real human beings, creatures inescapably bound to the contingent, empirically discoverable features of our biology, psychology and sociology.

Rather than resembling the comprehensive justification Enlightenment rationalists offered for their political standards, a full philosophical justification of political standards can instead resemble the justification offered by Enlightenment sentimentalists. Hume, Smith and Herder did not search for a normative vindication of their conceptions of justice grounded in

necessary principles of pure reason, but neither were they satisfied to see their commitments merely as the result of compromise among the myriad worldviews inevitably present in the modern nation-state. Instead, their moral and political commitments are both motivated by, and justified with reference to, empirical features of the human psyche. In this way, Enlightenment sentimentalism can join Enlightenment rationalism as one of many comprehensive doctrines participating in the overlapping consensus described in Rawls's later work.

What is more, the sentimentalist approach to political theory seems particularly suited to the modern, diverse societies which Rawls discusses. Although citizens in such societies may have many opposing moral convictions, they must come to some agreement on basic principles of justice. While some might hope to build a consensus behind these principles on the basis of reason alone, this is not the only faculty that all of us share. Given that our task is to build a just society for human beings, and not for rational beings as such, there is no reason why we cannot also appeal to the many non-rational features of the human psyche which we possess in common—our emotion, our imagination, and our ability to share in the inner life of others via sympathy. To forego these rich resources in either political theory or political practice would be a terrible waste. Although Rawls himself exaggerates the extent of his debts to Kant and minimizes the extent of his debts to Hume and Smith, much of the richness of his own political theory stems from the fact that it implicitly incorporates the insights of the sentimentalist as well as the rationalist Enlightenment.

This essay draws on work first presented as part of a doctoral dissertation defended at Princeton University in May 2006. I would like to thank my dissertation committee—Stephen Macedo, Jeffrey Stout, Charles Beitz, and Allen Patten—for their invaluable advice. I would also like to thank Eric Beerbohm, Corey Brettschneider, Mary Dietz, Coral Frazer, Martha Frazer, Katie Gallagher, John Holzwarth, George Kateb, Sharon Krause, Jack Turner, Alex Zakaras and two anonymous reviewers for their assistance. An earlier version of this essay was presented at the 2006 Annual Meeting of the Association for Political Theory, and I would like to thank my fellow panelists— Detlef von Daniels, Elizabeth Ellis, Mika Lavaque-Manty, and J. Donald Moon—for their participation.

[1] See John Rawls, *A Theory of Justice*. Revised Edition. Cambridge, MA: Harvard University Press, 1971/1999, especially pp. 40-46. (Henceforth cited parenthetically as *TJ*.)

[2] It is important to note that this political metaphor took a religious detour when, in pre-Enlightenment moral philosophy, God was seen as the authoritative legislator of moral standards. The Enlightenment notion of reflective moral autonomy developed more as a reaction to this theistic conception of moral legislation than as a direct application of political ideas to moral philosophy. For a thorough history of this development, see J. B. Schneewind, *The Invention of Autonomy: A History of Modern Moral Philosophy*. New York: Cambridge University Press, 1998.

[3] The distinction between a "Humean" and a "Kantian" or a "sentimentalist" and a "rationalist" approach to ethical reflection is commonplace among moral philosophers, intellectual historians and political theorists. I prefer the latter set of labels to the former in order to distinguish broader

Enlightenment-era intellectual currents from the work of any particular authors. Moral-philosophical sentimentalism in this sense must be distinguished from the concurrent movement of literary sentimentalism, associated with such emotionally overwrought novels as Jean-Jacques Rousseau's *The New Heloise,* Laurence Sterne's *Sentimental Journey*, Henry Mackenzie's *The Man of Feeling* and J. W. Goethe's *The Sorrows of Young Werther.* It was literary, not moral-philosophical, sentimentalism which gave the term "sentimentalist" the unfortunately mawkish connotations that it carries today. It is also important not to confuse the distinction between the rationalist and sentimentalist Enlightenments with the distinctions that have been drawn among the various "national" Enlightenments. Admittedly, many of the greatest thinkers of the sentimentalist Enlightenment—such as Anthony Ashley Cooper of Shaftesbury, Joseph Butler, Francis Hutcheson, David Hume and Adam Smith—were English or Scottish, while many of the greatest rationalists of the period were French or German. Yet there were many rationalists in Britain—among them Samuel Clarke and William Wollaston—just as there were many sentimentalists on the continent.

[4] "The moral sentiments arguments of the Scottish Enlightenment thinkers," Joan C. Tronto observes, "represent the 'losing' side in moral thinking in the eighteenth century" (Tronto, *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge, 1993, p. 36).

[5] Although both Enlightenments share a commitment to this ideal, the term "reflective autonomy" is my own, and was not used in the eighteenth century. My name for this shared ideal intentionally combines terms from the two Enlightenments. It was the sentimentalists who spoke most often of "reflection" and of humans as "reflective" beings; rationalists of course preferred to speak of humans as "rational" beings. And it was the rationalists who most often spoke of "autonomy." Yet the sentimentalists clearly saw the reflection they describe as autonomous,

while the rationalists clearly saw the autonomy they describe as reflective, even as each avoided the other's terminology.

[6] David Hume, *A Treatise of Human Nature*, 1739-1740. Edited by David Fate Norton and Mary J. Norton. New York: Oxford University Press, 2000, 2.3.3.4, p. 266.

[7] Ibid., 3.2.2.14, p. 317. This psychological holism is common to most Enlightenment sentimentalists, and is especially prominent in Herder's work. "The thought processes of our mind are undivided entities," Herder writes, "producing in their totality the diverse effects or manifestations which we treat as separate faculties" (*J. G. Herder on Social and Political Culture*. Translated and Edited by F. M Barnard. New York: Cambridge University Press, 1969, p. 259). If we sometimes speak the faculties of the human mind as separate entities, Herder explains, it is only as a philosophical abstraction, "because our weak spirit was unable to consider them all at once." (*Herder: Philosophical Writings*. Translated and Edited by Michael N. Forster. New York: Cambridge University Press, 2002, p. 83).

[8] For a fuller analysis of Kant's conception of the self along these lines, see Philip Fisher, *The Vehement Passions*. Princeton, NJ: Princeton University Press, 2002, pp. 234-236. For Rawls's own interpretation of Kant on the autonomy of reason and the heteronomy of sentiment, see Rawls, *Lectures on the History of Moral Philosophy*. Edited by Barbara Herman. Cambridge, MA: Harvard University Press, 2000, pp. 226-230 and pp. 280-285.

[9] Although Kant has been seen as unduly opposed to human emotion since at least Schiller, an important stream in recent Kant scholarship has sought to correct exaggerations of his actual position. For a summary of this literature, see Nancy Sherman, *Making a Necessity of Virtue: Aristotle and Kant on Virtue*. New York: Cambridge University Press, 1997, p. 4. See also Susan

Meld Shell, *The Embodiment of Reason: Kant on Spirit, Generation and Community*. Chicago: The University of Chicago Press, 1996.

[10] "Considered in themselves," Kant writes, "natural inclinations are *good*, i.e., not reprehensible, and to want to extirpate them would not only be futile but blameworthy as well; we must rather only curb them" (*Religion within the Boundaries of Mere Reason* 6:58, in *Religion and Rational Theology.* Translated and Edited by Allen W. Wood George di Giovanni. New York: Cambridge University Press, 1996, p. 101). A feeling or inclination is objectionable only when it "precedes consideration of what is duty and becomes the determining ground" of our action. When an inclination cannot keep to its proper, subservient place in the soul, it becomes "burdensome to right-thinking persons, brings their considered maxims into confusion, and produces the wish to be freed from it and subject to lawgiving reason alone" (*Critique of Practical Reason* 5:117-118, in *Practical Philosophy*. Translated and Edited by Mary J. Gregor. New York: Cambridge University Press, 1996, p. 235). For Rawls's interpretation of Kant on "the supremacy of reason," see Rawls, *Lectures on the History of Moral Philosophy*, op. cit. , pp. 200-207, p. 224.

[11] See Robert B. Louden, *Kant's Impure Ethics: From Rational Beings to Human Beings*. New York: Oxford University Press, 2000.

[12] "A mind will never be able to bear it own survey," Hume writes, "that has been wanting in its part to mankind and society" (Hume, *Treatise* 3.3.6.6, p. 395). This quotation is central to Annette Baier's interpretation of Hume, whose influence will be evident throughout this essay; see, for example, Annette C. Baier, *Progress of Sentiments: Reflections on Hume's Treatise.* Cambridge, MA: Harvard University Press, 1991, p. 96.

[13] Lawrence Kohlberg, *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. Volume II of *Essays on Moral Development.* San Francisco: Harper and Row, 1984, p. 173 (Henceforth cited parenthetically as *PMD)*, p. 194.

[14] I had been playing this game for years before I discovered that Annette Baier assigned it as a final exercise in her introductory ethics class. Baier, her students, and I all agreed in our classification of Hume as a Stage 3 subject (Annette C. Baier, *Moral Prejudices: Essays on Ethics*. Cambridge, MA: Harvard University Press, 1995).

[15] Kohlberg, *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*. Volume I of *Essays on Moral Development.* San Francisco: Harper and Row, 1981, p. 150.

[16] Carol Gilligan, *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press, 1982/1993, p. 18. (Henceforth cited parenthetically as *DV*).

[17] See Nel Noddings, *Caring: A Feminist Approach to Ethics and Moral Education*. Berkley, CA: University of California Press, 1984, p. 79, as well as Tronto, op. cit., p. 20

[18] Baier, 1995, p. 2. It is "an ironic historical detail," Baier writes, that Hume "showed less respect than we would have liked for those of his fellow persons who were most likely to find his moral theory in line with their own insights" (Ibid., p. 52).

[19] Reviewing the many studies on the subject which accumulated during the 1980's, Susan Moller Okin concludes that "the evidence for differences in women's and men's ways of thinking about moral issues is not (at least yet) very clear; neither is the evidence about the source of whatever differences there might be" (Susan Moller Okin, *Justice Gender and the Family*. New York: Basic Books, 1989, p. 15). For another review of the many studies giving us reason to doubt Gilligan's empirical findings, see Tronto, op. cit., pp. 82-85.

[20] This is approximately the role that Rawls plays in his appearances throughout Baier's *Moral Prejudices* (1995, op. cit.)

[21] See, for example, *TJ*, p. xvii, p. 20 (footnote) and p. 233. Rawls does acknowledge, however, that "The kind of utilitarianism espoused by Hume… is not strictly speaking utilitarian… All Hume seems to mean by utility is the general interests of necessities of society," in contrast to the pleasure-maximizing calculus of Bentham and later (proper) utilitarians (*TJ,* pp. 28-29). Although Smith's views are further still from those of classical utilitarianism, both Hume and Smith are presented as part of a continuous utilitarian tradition throughout Rawls, *Lectures on the History of Political Philosophy*. Edited by Samuel Freeman. Cambridge, MA: Harvard University Press, 2007. Hume's sentimentalism is addressed only in Rawls, *Lectures on the History of Moral Philosophy*, op. cit., pp. 21-102. Interestingly, Rawls gave extensive lectures on the sentimentalist Joseph Butler, whose anti-utilitarian but nonetheless sentimentalist ethics foreshadow those of Smith. See Rawls, *Lectures on the History of Political Philosophy*, pp. 416-457.

[22] Susan Moller Okin, "Reason and Feeling in Thinking about Justice," *Ethics* 99:2, 1989, pp. 229-249, p. 230. (Henceforth cited parenthetically as *RF*.) Some, but not all, of this article is incorporated into *Justice, Gender and the Family* (1989, op. cit.) in its chapter on Rawls (pp. 89-110). For a more recent account of reason and feeling in Rawls's moral psychology—inspired, like my own, by Okin—see Sharon Krause, "Desiring Justice: Motivation and Justification in Rawls and Habermas," *Contemporary Political Theory* 4 (2005), pp. 363-385, especially pp. 367-368.

[23] Okin, *Justice, Gender and the Family*, p. 15.

[24] On the falseness of the dichotomy between care and justice, see Michael Slote, *Morals from Motives*. New York: Oxford University Press, 2001, especially pp. 92-113; Eamonn Callan, *Creating Citizens: Public Education and Liberal Democracy.* New York: Oxford University Press, 1997, especially pp. 70-81 and Tronto, op. cit., pp. 166-167.

[25] Philip Fisher describes Rawls's original position as "expressly designed to let us imagine creating the social world dispassionately and impersonally," hence rendering *A Theory of Justice* "a book in which the passions play almost no part" (Fisher, p. 196). For another interpretation of the original position along these lines, see Cheryl Hall, *The Trouble with Passion: Political Theory Beyond the Reign of Reason*. New York: Routledge. 2005, pp. 31-35.

[26] It is important to distinguish here between two forms of rationalism: the ethical rationalism of Kantian moral philosophers and the amoral rationalism of rational-choice theorists. Although both sorts of rationalists agree that proper practical deliberation is the work of reason alone, Kantians argue that reason provides a categorical moral law commanding us to treat others as ends in themselves, while rational-choice theorists argue that reason instructs us only to pursue our individual self-interest. Rawls's greatest contribution to the rationalist tradition is to use the tools of the latter sort of rationalism in order to advance the cause of the former—crafting a thought experiment in which ignorance forces those rationally pursuing their individual self-interest to agree to moral principles which treat all as ends in themselves. Since both forms of rationalism reject the sentimentalist notion that proper reflection and deliberation involve the human mind as a whole, the complex relationship between the two is outside the scope of the present essay.

[27] "Empathy" was a term as yet unavailable to Hume and Smith, coined as it was by social psychologists early in the twentieth century from the Greek (*em-pathos*) as a rough translation of

Herder's German coinage *Einfühlung*, literally "feeling-into." "Sympathy," from the Greek for "feeling with" or "suffering with" (*sym-pathos)* was used by Hume and Smith in a sense much broader than that for which the term is normally used today, in a way that sometimes came closer to today's notion of "empathy." (For etymological information, see the unabridged Oxford English Dictionary, available online at http://dictionary.oed.com.)

[28] Elizabeth S. Anderson, "Is Women's Labor a Commodity?" *Philosophy and Public Affairs* 19:1 (1990), pp. 71-92, p. 81.

[29] See the discussion of Rawls's "morality of principles" as maintaining the affective concern for others first developed through the "morality of association" in Callan, op. cit., pp. 93-94. Although Callan rejects what he calls "sentimentality" (pp. 103-112), cultivating what can properly be called the sentimentalist features of Rawls's developmental moral psychology is nonetheless the focus of Callan's theory of civic education.

[30] Rawls is thus genuinely critical of what he calls the "Manichean" elements in Kant's moral psychology—the moments at which Kant suggests an ontological divide between the moral, rational and autonomous noumenal self and the amoral, emotional and heteronomous phenomenal self. Despite the fact that "the manner and tone of the Manichean tendency are often present" in Kant's writings, Rawls nonetheless insists that Kant's "explicit doctrine is Augustinian" (Rawls, *Lectures on the History of Moral Philosophy*, op. cit, p. 303). Rather than two sharply divided selves—one moral, one amoral—Rawls insists that Kant's moral psychology, charitably interpreted, involves the cultivation of a single, integrated self in which all of our faculties are harmoniously arranged for moral action. Of course, any plausible interpretation of Kant must still maintain that this integrated self is to be hierarchically organized with reason sovereign and inclination subject, but Rawls's charitable interpretation of Kant does

bring Kant's rationalist moral psychology slightly closer to Rawls's own implicitly sentimentalist position.

[31] For an excellent discussion of this important distinction, see Christine M. Korsgaard, *The Sources of Normativity*. With responses by G. A. Cohen, Raymond Geuss, Thomas Nagel and Bernard Williams. Edited by Onora O'Neill. New York: Cambridge University Press, 1996, pp. 8-10.

[32] Rawls, *Lectures on the History of Moral Philosophy,* op. cit., pp. 98-100.

[33] Korsgaard, p. 15. Korsgaard's example here is not of a Marxist, Freudian or Nietzschean theory of moral psychology, but of a Darwinian one. There is no reason, however, to think that believing our moral psychology has its origins in natural selection will prevent our reflective endorsement of the commitments to which this psychology leads us—unlike, say, believing that this psychology represents an internalization of the interests of a ruling class of capitalists or clergy. For purposes of her example, Korsgaard merely stipulates that the Darwinian theory does so, without claiming that this is actually the case. Indeed, she doubts that it is true even of Nietzschean and Freudian theories, which, rather than preventing the reflective endorsement of any moral commitments, merely believe "that our moral nature needs to be reformed and modified in certain ways in order to prevent it from making us ill" (p. 78, fn.).

[34] Ibid., p. 47, p. 49.

[35] Rawls, *Lectures*, 2000, p. 99. Rawls is of course paraphrasing the quotation from Hume's *Treatise* (3.3.6.6, p. 395) quoted earlier.

[36] Jeffrey Stout argues that there are in fact "two Rawlses" with regard to such issues of "metaphilosophy," only one of which is a Kantian rationalist. See Jeffrey Stout, *The Flight From*

Authority: Religion, Morality and the Quest for Autonomy. Notre Dame, IN: University of Notre Dame Press, 1981, pp. 222-223.

[37] That Rawls should have a basically cognitive theory of reflective equilibrium is unsurprising given that he borrows the concept from epistemology. Specifically, Rawls is inspired by Nelson Goodman's discussion of reflective equilibrium in Goodman's justification of the principles of deductive and inductive inference. See the citation of Goodman, *Fact, Fiction and Forecast*. Cambridge, MA: Harvard University Press, 1955, pp. 65-68 in *A Theory of Justice*, p. 18, fn. The idea that reflective equilibrium can be understood in a more "extended" sense to include some role for the emotions, albeit not as robust a role as Hume's sentimentalism would advocate, was first suggested by Henry S. Richardson in *Practical Reasoning About Final Ends*. New York: Oxford University Press, 1997, pp. 183-190. I am grateful to Sharon Krause for this reference to Richardson's work.

[38] Rawls, *Political Liberalism*. Revised Paperback Edition. New York: Columbia University Press, 1993/1996, p. 10. (Henceforth cited parenthetically as *PL.)* This turn of phrase is first used in Rawls's "Justice as Fairness: Political Not Metaphysical," *Philosophy and Public Affairs*. Vol. 14 (1985), pp. 232-252. Reprinted in *Collected Papers*. Edited by Samuel Freeman. Cambridge, MA: Harvard University Press, 1999, pp. 388-414, p. 388. I take this essay to mark Rawls's real break with his views of justification as presented in *A Theory of Justice*, and all references to Rawls's "recent" or "later" works imply books and essays published in 1985 or after.

[39] "The Idea of an Overlapping Consensus," in *Collected Papers*, p. 424.

[40] "The Domain of the Political and Overlapping Consensus," in *Collected Papers*, p. 483.

[41] Rawls, *Justice as Fairness: A Restatement*. Edited by Erin Kelly. Cambridge, MA: Harvard University Press, 2001, p. 7. See also Rawls, *Lectures on the History of Political Philosophy*, op cit., p. 54.

[42] Rawls does suggest, however, that the idea of the reasonable has an affinity with T. M. Scanlon's contractualist approach to ethics more generally. See *PL*, p. 49, footnote.

[43] "The Idea of an Overlapping Consensus," in *Collected Papers*, pp. 426-427.

[44] Jean Hampton, "Should Political Philosophy be Done Without Metaphysics?" *Ethics*. Vol. 99, No. 4 (July 1989), pp. 791-814, p. 809.

[45] Rawls, *Justice as Fairness*, p. 128.

[46] Thomas Nagel, *Equality and Partiality*. New York: Oxford University Press, 1991, p. 8, p. 33.

Michael Frazer is an assistant professor of Government and Social Studies at Harvard University. He is currently at work on two book projects, the first entitled *The Enlightenment of Sympathy: Justice and the Moral Sentiments in Eighteenth-Century Political Thought* and the second *At the Heretic's Deathbed: Adam Smith's Betrayal of David Hume and the Birth of Modern Conservatism*. After receiving his B.A. from Yale University and his M.A. and Ph.D. from Princeton University, Dr. Frazer was a postdoctoral research associate in the Political Theory Project at Brown University.